INSIGHT REVIEWS: THOUGHT, TOXICOLOGY, SYNTHESIS AND PSYCHIATRY

9 July 2009 www.nature.com/nature Vol 460 No 7252

nature

CHROMOSOME INSTABILITY

Extra centrosomes do battle in a tug-of-war

DEVIANT SUPERVIVAL
Breaking the redshift barrier

CARBON NANOTUBES
DNA sorts them out

CYSTIC FIBROSIS
The long wait for progress

NATURE.COM
Podcast glucose

Abstractions

SECOND AUTHOR

No one knows much about the Arctic Ocean during the Late Cretaceous, a period between 65 million and 99 million years ago when carbon dioxide levels in Earth's atmosphere were higher than they are today. Yet an understanding of conditions during that time might provide scientists with insight into what lies ahead, as CO_2 levels rise as a result of increasing greenhouse-gas emissions.

From analysis of today's oceans, scientists had thought that phytoplankton growth during that era was fuelled by upwelling — the wind-driven movement of nutrient-rich water towards the surface. But Alan Kemp, a palaeo-oceanographer from the UK National Oceanography Centre at the University of Southampton, and his colleagues have found that one of the main types of phytoplankton, diatom algae, became adapted to grow in different layers of ocean temperatures. This stratification let them bloom in spring and through summer when sunlight warmed the sea's upper layers (see page 254). Kemp tells *Nature* more.

Why is this finding important?

We know that the ocean is again becoming more stratified, as it was during the Late Cretaceous. Climatically, that period is an analogue for a warmer time with higher levels of CO_2, a likely scenario in the near future. This evidence of adaptability from the Cretaceous-era diatom fossils could be a clue to how an ocean may behave in the future, with more diatom algae species that can adapt to stratification. Upwelling-adapted diatoms tend to be small and reproduce rapidly, whereas stratified-adapted diatoms are usually larger and grow more slowly.

What does this mean for ocean ecosystems?

There could be a shift in the dominant or more successful species of plankton in the ocean, which is significant because the diatom algae are at the base of the whole food chain. If there are changes in the location of the dominant diatoms, including where they are most abundant, it could affect the fisheries and the entire ecosystem.

Did you go to the Arctic Ocean to get sediment samples?

No. In 1983, a team of Canadian scientists recovered shallow cores from the area's sea bed; incredibly, one contained diatom-rich sediments about 70 million years old. The cores were curated at the Bedford Institute of Oceanography in Nova Scotia, Canada, whose staff let us take samples. We used scanning electron microscopy to identify the types of diatom in these cores. Several genera and even some species from the Cretaceous are still present in the modern ocean, enabling us to make links with that period. ■

MAKING THE PAPER

Neil Ganem

A defect in dividing cancer cells results in a chromosome tug-of-war.

As Neil Ganem spent many hours peering down a microscope in search of dividing cells, his adviser David Pellman at the Dana-Farber Cancer Institute in Boston, Massachusetts, would sometimes joke that he was "looking at his wife's teeth".

Pellman was referring to philosopher Bertrand Russell's words about Aristotle's assertion that women have fewer teeth than men: "Although he was twice married, it never occurred to him to verify this statement by examining his wives' mouths." Ganem was testing, by direct observation, a long-held assumption in cell biology — and, similarly to Aristotle's assertion, it turned out to be wrong.

Most cancer cells have extra centrosomes — small structures that control the formation of the mitotic spindle, which organizes and segregates chromosomes during cell division. Dividing cells normally have two centrosomes, one at each end of the cell, which pull chromosomes towards the poles. If everything works, the cell splits into two daughter cells with equal complements of chromosomes. But many cancer cells are chromosomally unstable, meaning that they often 'missegregate' their chromosomes.

Many scientists assumed that the extra centrosomes in cancer cells generated this instability by giving rise to multipolar cell divisions, resulting in three or more viable daughter cells with abnormal chromosome numbers. "But no one had looked to see if it was really happening," says Ganem.

Using a microscope equipped with an incubator, Ganem watched thousands of cells grow and divide, looking for those with multiple poles and following the fate of their daughter cells — sometimes getting motion-sickness in the process, he laughs.

He found that cells with extra centrosomes rarely produce multipolar divisions. And when they do, the daughter cells typically perish. Having established that extra centrosomes do not lead to chromosome instability through multipolar division, Ganem asked: "Do extra centrosomes contribute to instability at all?"

To address this, he was inspired by the work of his graduate adviser, Duane Compton of Dartmouth Medical School in Hanover, New Hampshire. In normal cell division, one copy of each chromosome attaches to each pole. But sometimes one chromosome copy attaches to both poles — a defect known as merotely. "If not corrected, the chromosome can get stuck in a tug of war, and sometimes ends up going to the wrong daughter," explains Ganem.

Compton showed that merotely was a major cause of chromosome missegregation in cells with chromosome instability. So Ganem decided to test whether having extra centrosomes leads to merotely. He created two sets of cells, one set with extra centrosomes and the other with two centrosomes as normal. Turning to the microscope again, Ganem and colleagues found that cells with extra centrosomes were much more likely to form merotelic attachments and missegregate their chromosomes (see page 278). "Having extra centrosomes is definitely a defect that promotes missegregation," says Ganem.

The finding raises several questions to follow up, such as where the extra centrosomes come from. But for now, Ganem, who undertook the study as a side project, will be turning his attention back to his main research — a genome-wide screen to identify novel tumour suppressors. ■

FROM THE BLOGOSPHERE

The Great Beyond blog recounts how the blogosphere and Internet search engines are being used to debate scientific claims. Reporter Lucas Laursen has been following the activities of several bloggers commenting on the case of UK science writer Simon Singh, who was sued by the British Chiropractic Association (BCA) for libel (http://tinyurl.com/m43bf3).

In *The Guardian* newspaper last year, Singh questioned the scientific validity of using chiropractic techniques to treat children with ailments such as colic and asthma, and chided BCA members for promoting the treatments. (For more on the case, see http://tinyurl.com/m4ejy2 and http://tinyurl.com/maruzn.)

The BCA has now released a list of 29 studies that it says "support the claims which Dr. Singh stated were bogus". But several bloggers, such as Martin Robbins on Lay Scientist (http://tinyurl.com/kj7268), have pointed out some of the list's failings. Laursen notes that some studies "failed to conform to the statistically powerful, randomised, placebo-controlled, double-blind standard to which many medical studies are subject". ■

Visit Nautilus for regular news relevant to *Nature* authors ▸ http://blogs.nature.com/nautilus and see Peer-to-Peer for news for peer reviewers and about peer review ▸ http://blogs.nature.com/peer-to-peer.

nature

www.nature.com/nature

Vol 460 | Issue no. 7252 | 9 July 2009

Japan's tipping point

With changing demographics, a tight economy and increasing competition, Japan could slide from the top ranks of research nations. Drastic action is needed.

In May last year, Japan's science and education ministry issued a white paper that sounded an alarm for science and technology: because of increasing international competition, especially from China, and a declining young work force at home, Japan might "be forced out of the ranks of major international players and will risk losing [its] current plentiful and stable lifestyle".

Policy-makers have long worked to keep Japan in the top class. Last October, in a timely reminder of Japan's potential, four scientists who were born and educated in Japan received Nobel prizes. But can Japan continue to produce large quantities of high-calibre science as its researchers age and science-hungry university students grow scarce?

The outlook is grim, if this year's white paper, released last month, is anything to go by. Between 1998 and 2007, the number of researchers in universities rose by 15%, from 146,000 to 168,000. But in the same period, the number of researchers younger than 37 years old shrank from 36,773 to 35,788, and now only account for 21% of the total. The future is bleaker. The number of university students who want to study science and engineering plummeted from roughly 1 million in 1992 to around 630,000 in 2008.

Like its predecessors, the new white paper focuses on the possibility of invigorating the research community by increasing international exchange, recruiting women and encouraging young Japanese scientists. It should be possible to overcome Japan's famous cultural and language barriers and increase the flux of researchers across its borders. The RIKEN institutes do a pretty good job, and RIKEN's Functional Annotation of the Mouse cDNA project has been a shining example of how to coordinate an international project that continues to attract top foreigners to Japan.

Foreign investment

But that challenge is huge. After decades of extolling internationalization, only 10% of PhDs from Japanese universities go to people from other countries (compared with 42% in the United States and 41% in the United Kingdom). The percentage of foreign scientists at the country's universities and research institutes sits at a meagre 1.34%.

Alarmingly, Japanese scientists are less keen to go abroad, forgoing the international experience that produced some of Japan's most creative minds. Although two of the four Nobel prizewinners had worked with US institutes, the number of researchers spending more than 3 months in a foreign laboratory has dropped from 7,118 in 1997 to 4,163 in 2006. According to last month's white paper, only 2% of Japanese researchers had plans to work overseas. As international competition for scientific talent intensifies, Japan is closing in on itself.

Meanwhile, the struggle to encourage women in science, after years of effort, has produced only modest gains. The 12% of researcher posts held by women still places Japan near the bottom of the heap. Attitudes change slowly, and Japan is still a long way from fully tapping this resource.

Japan puts huge effort — at least on paper — into programmes to encourage young scientists. At least six ministries have such schemes. For instance, under the encouragement of the science and education ministry, 28 institutions have developed tenure-track systems that offer independence to younger scientists. And competitive grants for young scientists now contain about 30% towards indirect costs, which gives them more flexibility.

These plans are a promising start but it will take time to determine whether these movements will be able to level the top-heavy, deference-laden research structure currently in place. It doesn't help that young researchers will face a tightening job market as the science ministry reduces the money allocated to universities by 1% per year.

The country needs to experiment with aggressive policies. Initiatives such as the Kyoto University International Young Scientists Career Development Organization, which deliberately seeks to have one-third of the positions for young scientists filled by women and half by foreigners, should be duplicated and supported elsewhere.

"How much longer can Japan afford to lose the talent that its system is either chewing up or simply not developing properly?"

Japan's dependence on domestic scientists and its need for fresh ideas also justifies another radical investment in its longer term future: a ¥9.5-billion (US$100-million) science high school in Yokohama (see page 171). Other metropolitan centres should consider the benefits of such investment.

Overall Japan's science budget — at 3.61% of its gross domestic product — puts Japan near the top of the expenditure league tables. But the share of government-funded research is low. A ¥270-billion gift, courtesy of the government's economic stimulus packages, will fund 30 projects over 5 years. This is a huge amount of money. The package is likely to boost applied research, but in the interests of Japan's longer-term research fitness, such a sum should be spent extending the life of competitive grants, supplying more of them and creating new, independent tenure-track positions for young researchers at universities.

With little hope for a massive influx of creative thinkers from outside, Japan needs to fix the system that frowns on giving professorships and other opportunities to young independent scientists. How much longer can Japan afford to lose the talent that its system is either chewing up or simply not developing properly?

In its chronic failure to provide sufficient incentives and support for young researchers' independence, Japan as a scientific power is marching right past its tipping point. ■

How to stop blogging

Organizers have only two options for their meetings: open or closed.

Is the scientific conference in its death throes? Researchers have long anguished about the hyper-competitive culture that leads attendees to suppress their most interesting unpublished results. Such protectiveness can only be worsened by the increasing dissemination of results beyond the conference hall by bloggers.

Those who attempt to regain control of communications face outcry. Certain corners of the Internet have been erupting in argument in the past weeks following an announcement by Cold Spring Harbor Laboratory in New York that it will henceforth require scientists who blog to ask the permission of presenters before firing up computers or mobile phones and publicizing their findings.

We are in the midst of a clash of conference-going cultures (see *Nature* **459,** 1050–1051; 2009). Attendees who have taken to blogs and other social-media applications such as Twitter and Friend Feed will value the instantaneous communication of fact, conjecture and commentary as a way to network beyond badge-holders. Most researchers, in contrast, will focus on the science and ways to network with fellow attendees. If they are aware of social-networking applications, they are likely to regard them as distractions at best. At worst, they will fear them as tools to undermine and scoop, to release data not ready for consumption by anyone other than the trusted colleagues who bothered to make it to their talk or walk up to their poster and start asking questions.

Conference organizers are stuck in the middle. They want to let the world know that their meetings are worthwhile, and yet they also want to attract speakers presenting the newest and most cutting-edge findings. So how to protect speakers from having sensitive, unfinished or 'scoopable' work broadcast to the world?

Cold Spring Harbor's strategy is to unilaterally require all bloggers to ask permission: an opt-in strategy for presenters. Others have proposed an opt-out strategy: speakers can affix a logo to their slides or posters that says, in effect, "These results are off limits, for now."

Neither of theses solutions is sustainable, however. Anyone who's heard the chime of a digital camera starting up in the middle of a session knows that clear, sometimes quite threatening, restrictions on photography are regularly ignored. So too it will be with social media users, for whom jotting down their personal reactions to a talk or poster on the fly is simply second nature.

So, what is the answer? *Nature* has made the case that blogging by researchers is good. Critical discussion of worthy results should not in principle be restricted to walls of a conference hall or even the pages of a journal. Any meeting to which anyone can register is fair game for all available communications technologies — and any rules that cannot be policed will be ignored anyway.

"Anyone who's heard the chime of a digital camera in the middle of a session knows that restrictions are regularly ignored."

But some of the most valuable scientific meetings are more focused and deliberative, and are by invitation only. These intellectually intimate gatherings are valuable precisely because the researchers who attend feel free to speculate and to stimulate their colleagues, and to try to establish new research agendas. Whether an attendee is a blogger or a reporter or a practising researcher is immaterial. The meeting is 'off the record', and all have to sign up to that.

Closed or open? Meeting organizers need to be clear in their minds which of these two approaches is appropriate, and be explicit about it from start to finish. The consequence that, in competitive fields, presentations at open meetings will become even more protective and boring is an inevitable consequence of the Internet. ■

Adieu to nuclear recycling

President Barack Obama should be applauded for his decision to scrap commercial reprocessing.

This week, US President Barack Obama has been grabbing headlines with his efforts to revitalize the Strategic Arms Reduction Treaty — a US/Russian agreement to reduce the nuclear arsenals of both nations.

Such efforts will be applauded worldwide, but another decision by the Obama administration deserves equal acclaim. On 29 June, the president quietly cancelled a lengthy environmental review that was the first step in allowing the resumption of commercial nuclear reprocessing in the United States. Nuclear reprocessing chemically separates uranium and plutonium from spent nuclear fuel so that it can be reused in specialized reactors. The same technique can be used to purify material for nuclear weapons, and it is partly for that reason that the United States decided to halt reprocessing in the 1970s.

Obama's predecessor, George W. Bush, sought to reverse that decision. He thought that reprocessing could be part of a broader approach that would see used fuel from non-nuclear-weapons states brought to the United States for reprocessing. As part of the Global Nuclear Energy Partnership programme, Bush advocated the construction of a demonstration commercial reprocessing plant, and an environmental review was already under way when Obama came into office.

Such a plant, had the plans been allowed to continue, would have been both costly and counterproductive. Proliferation worries aside, reprocessing is complex, expensive and creates a liquefied stream of highly radioactive waste that is difficult to dispose of. The technology is likely to be needed within the next two decades, so Obama is right in his decision to allow research into ways to improve reprocessing, while constraining the programme to one of basic science.

The decision to halt commercial nuclear recycling sends a clear message that the United States is committed to nuclear non-proliferation. Such decisions, together with diplomacy such as that taking place in Russia, are deliberate and encouraging first steps towards building an international consensus on reducing the threat from nuclear weapons. ■

RESEARCH HIGHLIGHTS

ECOLOGY

A dusting of snow

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0900758106 (2009)

Dust deposited on mountain snow increases the amount of solar radiation absorbed, causing it to melt about a month earlier than clean white snow. Many ecologists have suspected that this leads to much earlier growth in previously snow-covered vegetation.

A team led by Heidi Steltzer at Colorado State University in Fort Collins set up 13 experimental plots in Colorado's San Juan Mountains, adding dust to some, scraping others clean, and covering some in dark fabric to control the timing of the thaw.

Contrary to expectation, the team found that after an early thaw — when temperatures were still freezing — vegetation didn't grow for more than two weeks. This delay, Steltzer says, might mean that nitrogen released in the melt could go to waste and may contribute to the elevated nitrogen levels seen in some alpine lakes.

EVOLUTION

Mary had a littler lamb

Science doi:10.1126/science.1173668 (2009)

Changes in a given species may be driven by evolution or by ecological changes. Discriminating between the two is fraught with difficulty.

Tim Coulson of Imperial College London and his colleagues say they have managed to tease out the different drivers for one such change. Soay sheep on the Scottish isle of Hirta, UK, have shrunk in size over the past 20 years, despite heavy sheep being more likely to survive when young.

In analysing the different contributions to this size decrease, the researchers found that climate and population density explained most of the change, with natural selection underlying a smaller portion.

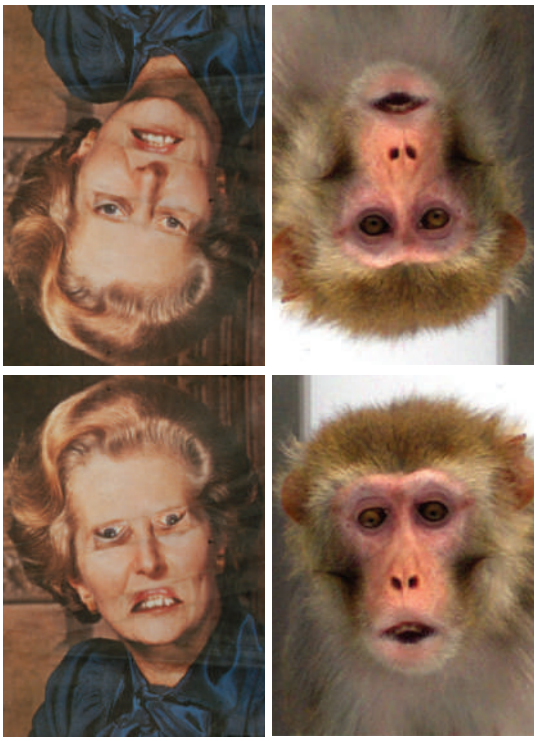
A. OZGUL/AAAS

Picture imperfect?

Curr. Biol. doi:10.1016/j.cub.2009.05.067 (2009)

The human talent for recognizing differences in faces relies on how facial features are configured. But flip an image of a face, and alterations as drastic as inverted mouths and eyes aren't as noticeable — a phenomenon known as the Thatcher effect. Robert Hampton of Emory University in Atlanta, Georgia, and his colleagues recently demonstrated that the effect is present in another primate.

They monitored the length of time rhesus monkeys (*Macaca mulatta*) looked at pictures of monkey faces. Over time, the animals became less interested in all images, but they spent significantly more time looking at the strange, upright altered (Thatcherized) photos than they did looking at the same images upside down.



LEFT: P. THOMPSON *PERCEPTION* **9**, 383–384 (1980)/PION, LONDON; RIGHT: B. BASILE/YERKES NATL PRIMATE RES. CENTER

CANCER BIOLOGY

At rest in the bones

Cancer Cell **16**, 67–78 (2009)

A cancer-promoting protein called Src helps breast-cancer cells to survive — sometimes for years — in the bone marrow.

Joan Massagué of the Memorial Sloan-Kettering Cancer Center in New York and his colleagues mined gene-expression data for 615 breast tumours. They found that cancers that relapsed five years or more after the initial cancer diagnosis bore a unique pattern of Src-regulated gene expression.

Reducing the amount of Src protein in human breast-cancer cells slows the outgrowth of these cells in bone in mouse models, but does not affect the growth of metastases in the lungs. Furthermore, Src-depleted cells are unable to suppress a cell-death pathway that is expressed in the bone marrow and so die off.

ASTRONOMY

A star is born

Astrophys. J. **699**, 1300–1306 (2009)

The birth of young stars is shrouded in mystery because they tend to form at the centre of giant clouds of dust and hydrogen gas. Only light in the millimetre- and infrared-wavelength range can pass through the clouds, and astronomers have historically lacked telescopes that are capable of clearly discerning those wavelengths.

Now, Jonathan Williams of the University of Hawaii and his colleagues have used the latest generation of telescopes to capture a high-resolution view of a nearby star-forming cluster. They found five objects in the cluster, including a bright protostar and a starless, collapsing core of gas. The observations show that a single cluster can give birth to a plethora of different protostellar types, and may help to give astronomers a better understanding of the birthing process.

GENOMICS

Closing in on cholesterol

Cell Metab. **10**, 63–75 (2009)

In an attempt to uncover genes that regulate cholesterol levels in cells, a team of researchers has used a bevy of screening approaches and identified transmembrane protein 97 (TMEM97) as a new candidate for cholesterol regulation.

The researchers, led by Rainer Pepperkok and Heiko Runz of the European Molecular Biology Laboratory in Heidelberg, Germany, and the medical faculty of the University of Heidelberg, depleted cholesterol from cells and looked at the response in gene

expression. They then performed literature searches and used RNA-silencing techniques to further narrow down their results.

The team focused on TMEM97 because it is highly expressed in sterol-depleted cells and is targeted by sterol-response-element binding proteins, which also activate other cholesterol regulatory genes. Knocking down its expression led to decreased cholesterol uptake and lowered cholesterol levels in cells. The authors suggest that their combined screening technique offers a powerful method for pinpointing genes in particular pathways.

MICROBIOLOGY

Supershedding mice

Infect. Immun. doi:10.1128/IAI.00558-09 (2009)

Clostridium difficile, a spore-forming anaerobic bacterium that inhabits the gut of as many as 5% of humans, is a leading cause of antibiotic-induced diarrhoea. When a carrier takes antibiotics, these can wipe out other gut microbiota and send *C. difficile* into overdrive, which can lead to rampant spread of the bacterium in hospital settings. Trevor Lawley of the Wellcome Trust Sanger Institute in Hinxton, UK, and his colleagues have developed a new mouse model for the condition.

Between two and three days after treatment with the antibiotic clindamycin, mice carrying *C. difficile* shed a million-fold more spores in their faeces than before. More than half of the mice retained this 'supershedder' status for several weeks. Uninfected mice housed with supershedders became carriers themselves; because the bacterial spores are ethanol resistant, cages had to be thoroughly cleaned with sporicides.

GEOLOGY

Earth-shattering research

Geology **37,** 615–618 (2009)

How does a rock weather away into clay? Usually, it is helped along by fungi, in a process now examined in detail by Steeve Bonneville at the University of Leeds, UK, and his colleagues.

The team grew tree seedlings and a symbiotic fungus *Paxillus involutus* in an experimental set-up that included the mineral biotite, and followed the action with high-resolution microscopy. The fungus enhances the weathering of biotite, not by exuding chemically active substances as was previously thought, but by physically disrupting the structure of the mineral with its hyphae. This is followed by chemical changes as the fungus removes potassium and oxidizes iron within the mineral. Soon, the rock is no more.

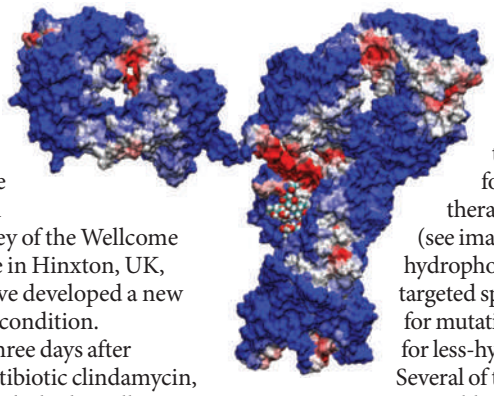
COMPUTATIONAL BIOLOGY

Unstuck by design

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0904191106 (2009)

Computational predictions could make it easier to design biopharmaceuticals that don't stick together. Aggregation can destroy the effectiveness of antibody-based drugs or, worse, spur an unwanted immune response. But current methods to stabilize antibodies involve trial and error, which can be costly.

Bernhardt Trout of the Massachusetts Institute of Technology in Cambridge and his colleagues have developed a method to predict the exposed areas on a protein that are hydrophobic and make it more likely to clump in solution. The researchers made these predictions for two different therapeutic antibodies (see image left; red represents hydrophobic patches) and targeted specific amino acids for mutation, swapping them for less-hydrophobic versions. Several of their creations were more stable than the originals.



NATL ACAD. SCIENCES

NEUROLOGY

Feeling out autism

Nature Neurosci. doi:10.1038/nn.2356 (2009)

Children with autism spectrum disorder (ASD) may rely more on physical feedback than visual cues when performing tasks.

Reza Shadmehr at Johns Hopkins University in Baltimore, Maryland, and his colleagues asked children with ASD and children with typical development to play a video game in which they learn to use a robotic arm to capture animals. As training progresses, a force is applied to the arm that the children must correct for. In some instances, the corrective force that the children apply was measured, revealing how their brains build a generalized mental model of the way in which the robotic arm works.

Those children with ASD who were most reliant on the physical feedback to build this mental model had larger deficits in motor control, imitation and even social function.

JOURNAL CLUB

James J. Collins
Howard Hughes Medical Institute, Boston University, Massachusetts

A bioengineer gets schooled by *Escherichia coli*.

The ability to learn from situations and to predict certain outcomes sets us apart from many living things. It prevents many of us from chasing balls into busy streets or placing bottles of ethanol near Bunsen burners. Still, it didn't stop thousands of US researchers submitting applications for the National Institutes of Health's Challenge Grants — funded by President Barack Obama's economic stimulus package — despite an expected success rate little better than one or two per cent.

To enjoy the benefits of learning and predictive behaviour, we usually assume that you need a nervous system or at least a neuron. So it was surprising to read that Saeed Tavazoie at Princeton University, New Jersey, and his colleagues have demonstrated that bacteria can learn and exhibit anticipatory behaviour (I. Tagkopoulos *et al. Science* **320,** 1313–1317; 2008). They show computationally and experimentally that *Escherichia coli* can learn temporal correlations between environmental stimuli — for example, that an increase in temperature is followed by a decrease in oxygen levels — allowing the bacteria to predict and prepare for future environmental changes.

The researchers show that this associative learning is accomplished by rewiring of biochemical networks. Strikingly, they also show that, like many of us, *E. coli* quickly 'unlearn' (in fewer than 100 generations) what they had learned in a new situation.

Now we know that bacteria can be taught such tricks, it will be interesting to see if we can use novel combinations of environmental stimuli to train microbes to efficiently convert biomass into energy sources, such as hydrogen or butanol. By providing *E. coli* with such an educational stimulus package, we may be able to boost the global economy.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

US stem-cell research expands

Nearly 11 years after Wisconsin-led researchers reported the first isolation of human embryonic stem cells, the field became eligible this week for broad research funding by the US government.

In final guidelines that went into effect on 7 July, the National Institutes of Health (NIH) established a process that will allow scientists who hold stem-cell lines derived before this date to apply for their inclusion in an agency-established registry of fundable cell lines.

"The NIH has done what is best for the field by having their own registry — one list that everyone can work from."

In doing so the agency, based in Bethesda, Maryland, acknowledged that the strict ethical conditions it laid out in draft guidelines published in April (see *Nature* **458**, 950–951; 2009) may have excluded many existing cell lines from federal funding. After receiving more than 49,000 comments on the draft guidelines, "it became clear that there were lines that responsible people would all agree were responsibly derived … but might not meet the exact guidelines that we have put forth," said Raynard Kington, acting NIH director, on 6 July.

A working group will be established to judge on a case-by-case basis whether pre-existing lines "meet the principles underlying the guidelines", he said.

Those principles remain unchanged from the April draft: fundable lines must be derived from embryos that were created solely for reproductive purposes and are no longer needed. Parents must voluntarily donate the embryos without inducements or researcher influence, and written informed consent must be obtained.

The final guidelines, like the draft ones, exclude funding for stem cells derived from embryos created for research, whether by *in vitro* fertilization, somatic cell nuclear transfer or parthenogenesis, when an unfertilized egg is developed into an embryo.

Although lines created in the United States from 7 July onwards will have to follow the guidelines to the letter, scientists who create lines abroad in future may apply to the working group to demonstrate that a line was made under procedural standards that are "at least equivalent" to those provided in the final guidelines. The NIH director will give the final approval to all decisions made by the group.

Kington said he expected the group to comprise nine to ten people, including ethicists, scientists and fertility doctors, and to be up and running "within the next couple of months", along with the new registry.

Many scientists were delighted with the final guidelines. "It's a huge step forward,"

Flu jabs urged for developing countries

SIENA

Influenza experts are recommending an extensive vaccination programme against seasonal flu in developing countries, in part to boost demand for vaccines so that firms can ramp up production to cope with pandemics. The message came from scientists and policy-makers who met on 2–3 July in Siena, Italy, to assess the gaps in their knowledge about the current H1N1 pandemic virus.

The governments of many developing countries remain to be convinced that flu is a major danger for their citizens relative to other health problems, says Abdullah Brooks of the International Centre for Diarrhoeal Disease Research in Dhaka, Bangladesh. Yet Brooks presented research showing that around one-third of pneumonia deaths in children younger than 2 years old in his region can be attributed to the influenza virus. According to the United Nations agency UNICEF, pneumonia kills more than 2 million children under the age of five each year — more than any other disease.

At the meeting, experts recommended that pilot studies be conducted in developing countries to measure the prevalence of flu virus in sick children, and to assess how much a flu-vaccination programme would reduce the burden of disease in the countries. UNICEF, health charities and the governments of rich nations would probably be approached for financial support.

As well as providing a major public-health benefit, the effort could create a larger, more stable market for seasonal flu vaccines in the future. "A few months ago we were discussing whether we would need to close some of our manufacturing plants because we were losing so much money on flu vaccines," says Rino Rappuoli, head of vaccine research at Novartis in Siena, adding that the current H1N1 pandemic has helped to avert any closures as governments race to stock up on vaccines. For example, the firm was awarded US$289 million by the US Department of Health and Human Services (HHS) in Bethesda, Maryland, in May to produce H1N1 vaccine antigen as well as an adjuvant to amplify the immune response to the vaccine, thus reducing the amount of antigen needed in each shot and stretching manufacturing-plant capacity. Other vaccine companies, including GlaxoSmithKline, Sanofi Pasteur, CSL Biotherapies and MedImmune, will also benefit from $643 million in HHS orders.

Boosting demand could help vaccine manufacturers keep their facilities open.

NOVARTIS AG

P. GREB/USDA

PESTS TAKE ON GM COTTON
Bollworms overcome two toxins made by transgenic crops.
www.nature.com/news

says George Daley, a researcher at Children's Hospital Boston and the Harvard Stem Cell Institute in Cambridge, Massachusetts. "It's flexible and science friendly."

Sean Morrison, a stem-cell biologist at the University of Michigan in Ann Arbor, adds: "The NIH has done what is best for the field by having their own registry — one list that everyone can work from."

Some scientists, including Daley, said that they were disappointed with the exclusion of embryos derived for research purposes, but pointed out that the agency intends to revisit the guidelines as the science evolves.

The NIH guidelines depart in one significant way from existing National Academy of Sciences standards; they do not require consent from gamete donors — only from the couple seeking *in vitro* fertilization services.

The guidelines respond to an executive order issued in March by President Barack Obama (see *Nature* **458**, 130–131; 2009), who lifted restrictions imposed by President George W. Bush on 9 August 2001. ■

Meredith Wadman

The US Food and Drug Administration (FDA) has yet to approve any flu vaccines that contain an adjuvant, a cautious response to the possibility that adjuvants could trigger autoimmune disease on very rare occasions. Yet in 2000, the European Medicines Agency approved Novartis's seasonal flu vaccine Fluad, which contains MF59 — an adjuvant based on squalene, a complex hydrocarbon that is the biochemical precursor to steroids. Novartis says that it has sold 45 million vaccination shots containing MF59, and that no side effects have been reported other than occasional inflammation at the injection site. "It isn't really clear what additional safety data the FDA could be waiting for," says Peter Palese, a virologist at Mount Sinai School of Medicine in New York. And a study published earlier this year has shown that MF59 can broaden the effectiveness of an H5N1 vaccine to tackle several similar viral strains — an effect that could potentially be applied in H1N1 vaccines as well (G. Galli *et al. Proc. Natl Acad. Sci. USA* 106, 7962–7967; 2009).

The consensus of the meeting was that H1N1 vaccines, which could be given to billions of people this year, should contain an adjuvant. These vaccination programmes will provide a unique opportunity to identify any extremely rare side effects, says virologist Albert Osterhaus of the University of Rotterdam, the Netherlands, as long as they have "mechanisms to monitor all cases of autoimmune disease to see if the vaccine is really to blame." ■

Alison Abbott

Czech researchers angry over government changes

Scientists in the Czech Republic are up in arms over drastic changes in the national science-funding system that they say will damage basic research in the long term.

Starting next year, core funding for Czech universities and research institutes will be allocated according to rigorous metrics. Institutes can obtain 'points' for a variety of publications and for patents, but also for any software, methods, samples, prototype devices and 'validated technologies' developed in-house. The government plan, drawn up over the past three years with the goal of increasing the efficiency of research and development, was approved by the Czech cabinet on 29 June.

Jiří Drahoš fears that budgets will be cut still further.

But critics say that the prescribed funding formula gives a grossly distorted picture of the real assets of Czech science, and threatens to destroy its best parts. Ill-defined criteria of 'innovation' will downgrade the merit of detailed work on books and peer-reviewed papers, they argue, while rewarding impressive-sounding concepts and rushed patent applications.

"One general formula for all branches of science?" says Jan Konvalinka, a group leader at the Academy of Sciences of the Czech Republic's Institute of Organic Chemistry and Biochemistry in Prague. "That's almost alchemy."

The academy, which runs 54 basic-research institutes, will be hit badly. In 2010, it will lose 1 billion Czech koruny (US$54 million), or about 20% of its annual budget, as a result of the new scheme. Academy president Jiří Drahoš says that in future years the number could climb to 45% of the value of the current budget.

"If this new formula is strictly applied, we will enter into a liquidation spiral," he says. "At some point the academy just cannot work anymore."

At a crisis meeting last week, the academy's leadership decided to stop paying for repairs and procurement, including purchasing scientific equipment, starting in January. If budgets continue to decline, staff may have to be laid off, says Drahoš.

The reform comes from a council chaired by Jan Fischer, the Czech prime minister, and comprising representatives from universities, the academy and industry. Initially, applied and industry-oriented research were supposed to be strengthened without affecting basic research. Changes included reallocating science portfolios to the industry ministry and creating a new technology agency.

The government had said previously that public-science budgets would increase by 2 billion koruny per year, to 31 billion in 2012. But the economic crisis has thwarted the plan, and science expenditure is now frozen at 25 billion koruny per year.

The freeze follows several prosperous years for Czech science, which has experienced a notable upswing since the country entered the European Union (EU) in 2004: it now spends around 1.4% of its gross domestic product, up from 1.2% in 2003, on research and development. This is more than any other formerly socialist, eastern European country in the EU.

Critics say that the new system is a step backwards. "It started out as a good idea, but then it got screwed up in the hands of technocrats," says Pavel Jungwirth, a computational chemist at the Institute of Organic Chemistry and Biochemistry.

Jan Zrzavý, a biologist at the University of South Bohemia in České Budějovice, who was on the government council that drafted the reform, declined to comment, saying that transferring the discussion to an international forum would be "ill-calculated and potentially harmful". ■

Quirin Schiermeier

SPECIAL REPORT

Developing nations tackle climate

Emissions targets, clean-energy projects and calls for justice are multiplying, reports **Jeff Tollefson**.

Following up on a non-binding pledge to halve its emissions by mid-century, the Mexican government is finalizing regulations that would curb the country's projected emissions by more than 6% over the next three years.

The goal of the regulatory programme, driven by President Felipe Calderón, is to demonstrate actual emissions reductions — totalling 50 million tonnes — before his government leaves office in 2012. It is meant as a down payment toward Mexico's pledge of halving emissions by 2050, the most aggressive planned cut among developing nations so far.

The new programme, which is expected to be finalized this month, is "our way of proving that you can take the climate change issue seriously and still lay a sound foundation for development", says Fernando Tudela, who heads the Mexico climate delegation and helped broker the deal for a domestic programme.

With this promise, Mexico is challenging the idea that developing nations are sitting on their hands as greenhouse-gas emissions skyrocket. It isn't alone: in the run-up to the United Nations climate summit in Copenhagen in December, other developing nations have pledged their own dramatic reductions. Brazil has promised to reduce deforestation by 70% by 2017, and South Africa says it could level off emissions by 2025. Still others, including China and India, are pouring money into green-energy projects; even South Korea, which gets most of its energy from fossil fuels, plans to funnel some US$40 billion into clean energy over the next several years.

Whether developing countries deliver on these promises remains to be seen, particularly given their reluctance to accept mandatory emissions targets. But existing unilateral actions may have a larger impact than many realize. Existing commitments by Brazil, China and Mexico alone will do nearly as much to reduce greenhouse-gas emissions by next year as the European Union hopes to accomplish by 2020, according to an analysis by the Center for Clean Air Policy in Washington DC.

Still, rapid economic growth in places such as China and India is tending to swamp any gains, which means that cumulative emissions keep rising. Emissions from developing countries already surpass those of industrialized nations and are projected to rise twice as fast as those in developed countries over the next couple of decades (see graphic, below).

"We're beginning to see a little more appreciation about what's going on in developing countries," says Ned Helme, president of the Center for Clean Air Policy. "But it's a hard message to get through."

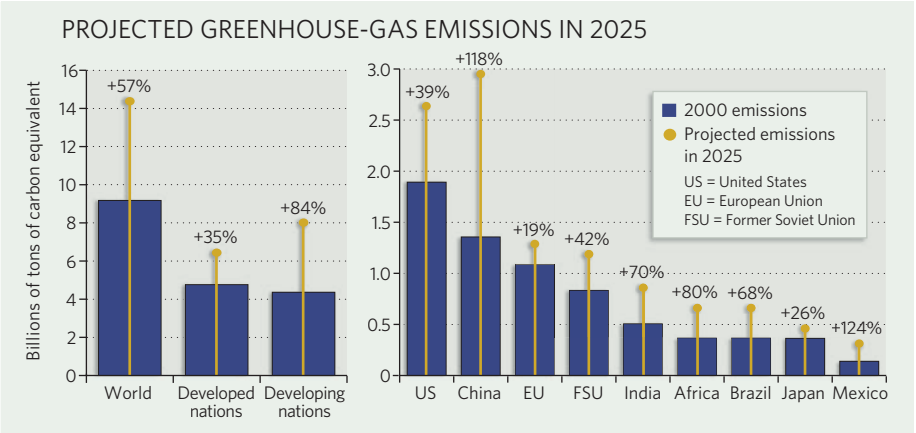
The gulf between developed and developing nations over greenhouse gases — and who should take responsibility for what — remains alarmingly wide. Many developing countries, including China, have argued that richer countries should curb emissions by 40% or more below 1990 levels by 2020. By contrast, the United States is talking about returning to roughly 1990 levels by 2020; even the European Union, which has the most aggressive targets of all the negotiating bodies at the UN talks, is proposing a cut of only 30% by 2020.

Meanwhile, developed and developing nations are at odds over what kind of commitments, if any, can be expected of the developing world. The dispute centres on four words — "common but differentiated responsibilities" — that are in the UN Framework Convention on Climate Change signed at the 1992 Earth Summit in Rio de Janeiro, Brazil. Those words were repeated in the 1997 Kyoto Protocol, which committed only developed nations to reducing their emissions and set up financial mechanisms to help pay for clean-energy and other projects in poor countries.

Proposals abound about how to structure a new deal in Copenhagen. But many of those being put forth by developing countries are based on historical emissions, as rich countries are mainly responsible for pollution dating back to the industrial revolution. Last month, at a UN climate meeting in Bonn, Germany, several countries, including Bolivia, Brazil, China and India, held a technical session on the issue of historical emissions and how they could be incorporated into the next climate treaty. "This is basically the fault line between developed and developing countries," says Lim Li Lin, a researcher with the Third World Network, a non-profit advocacy group based in Penang, Malaysia, which works with developing countries. She says proposals are increasingly linked to scientific assessments about what kind of emissions reductions are necessary, as opposed to what is politically possible. "We're coming to the discussion with a principle-based approach," she says.

China bases its assessment of historical emissions on the idea of a per-capita carbon budget for carbon emissions from 1850 to 2050, in which cumulative emissions, past and future, are allotted according to population. On this per-capita basis, emissions from developed countries — the 'Annex I' nations — are more than 10 times higher than Chinese emissions. If the goal were per-capita equity, Annex I countries would have had to stop emitting greenhouse gases entirely in 2007.

Bolivia carries this approach one step further by framing the issue in terms of 'climate debt'. Like China, it reckons that even if Annex I countries reduce their current carbon emissions they will still have exceeded their per-capita allocation. Everything that developed countries emit above their per-capita allotment incurs a climate debt that must be repaid with either money or technology. Bolivia has already attracted support from countries such as Malaysia, Paraguay and Venezuela, and



WHO

A CHEAPER RABIES VACCINE?
Single injection cures infected mice.
www.nature.com/news

R. WALLIS/CORBIS

India won't promise to cut its emissions in negotiations abroad but is building wind turbines at home.

more than 230 advocacy organizations, including the Third World Network, have endorsed the general idea.

One organization that has not backed Bolivia's proposal, or any other formula regarding historical emissions, is the Group of 77 (G77), the largest and most diverse coalition in the UN climate talks, which in spite of its name represents some 130 developing countries, including China. The G77 has called for a centralized fund to promote technology transfer, but its members are divided on key climate issues such as emissions restriction and adaptation funding.

The Alliance of Small Island States, for instance, often pushes for aggressive action to reduce greenhouse gases because of its pressing concern over rising sea levels, while those nations classed by the UN as Least Developed Countries take the lead on demanding funding for adaptation projects. Both groups clash with members of the Organization of the Petroleum Exporting Countries, who worry about falling oil demand as the world shifts towards renewable energy and who want international aid to diversify their economies.

Then there are major developing-world emitters such as Brazil, China, India and South Africa. These nations are big enough to get invited to the Major Economies Forum on Energy and Climate organized by the United States but frequently seek refuge within the G77, where expectations for commitments to mitigate climate change are much lower. "You have this huge issue of smaller countries that are in a way shielding larger-emitting countries from any kind of stronger commitment," says Ana Maria Kleymeyer, a former Argentinean negotiator who consults for G77 countries. But she says that dynamic is changing as Brazil, South Africa and even China hint that they might accept binding commitments, although they won't go as far as cutting emissions.

One country that continues to take a particularly hard line on emissions is India. Shyam Saran, Prime Minister Manmohan Singh's special envoy for climate change, says India needs money and technology before it will agree to any emissions reductions targets. Even then, he adds, agreeing to any kind of hard, quantifiable targets could be difficult. Indeed, Saran's only promise is that India won't exceed the per-capita emissions of the developed world — more a threat than a promise, given that the country has more than a billion people who each use relatively little energy today.

Domestically, however, India has developed an ambitious climate agenda, points out Rajendra Pachauri, chairman of the Intergovernmental Panel on Climate Change and a member of Singh's Council on Climate Change. The centrepiece of the plan, which is currently being finalized, is a programme that could produce upward of 20,000 megawatts of solar power by 2020, some 6–7% of the nation's projected power at that time, he says.

Pachauri acknowledges what he calls a "disconnect" between India's position on climate at home and abroad. He thinks this stems in part from delays to the climate agenda caused by unrelated distractions, including the terrorist attacks on Mumbai last November and national elections in May. But the government is now moving forward, he says. "I'm talking as an Indian and not as chairman of the IPCC, but I suspect by the time we go to Copenhagen, we will be more in line with our domestic and international positions."

In Mexico, climate experts are looking at ways to combine the country's domestic and international commitments. Its proposed programme would employ cap-and-trade to reduce emissions from private industries such as cement and paper, for example, but companies that curb pollution even further would also be able to access international funds.

Mexico has also proposed a $10-billion 'green fund' to help pay for things such as adaptation and technology transfer in the developing world. Under the plan, all nations would contribute to the fund based on their wealth and population; the money would be distributed to developing countries according to the same formula. The green-fund proposal tacitly acknowledges that 'developing' countries come in many flavours, and that many of them are in fact wealthier than 'developed' countries, as defined by the UN climate treaty. Mexican officials say the goal is to find a middle ground.

"We still keep the distinction, because I think it's real, between developed and developing," Tudela says. "But we begin to blur the boundaries, accepting a new way of differentiating the responsibilities." ■

For more climate coverage, see nature.com/roadtocopenhagen.

Q&A

Iran presidential candidate speaks out

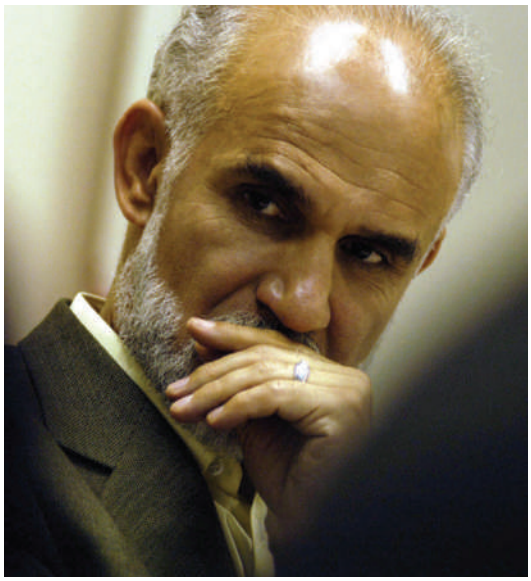
Biomedical researcher **Mostafa Moin** was a reformist candidate in Iran's 2005 presidential elections, following which Mahmoud Ahmadinejad became president. In a 2006 interview with *Nature*, Moin, a former minister for higher education and for science, argued that building a stronger civil and democratic society in Iran was key to the country's scientific development and it becoming a knowledge-based society. In an exclusive new interview, he discusses where the current situation leaves those ambitions.

What is your assessment of Mahmoud Ahmadinejad's track record on science, and on academic freedom and social reform?
Scientific activity has stagnated. Students and young scientists have been discouraged. Pressure has been exerted on scientific forums and centres, and academic freedom restricted. The brain drain of elites has worsened. Administrations have been massively populated with people from the military and the security forces. We have an economic recession; destruction of ethical norms and public culture; and the scientific, political and cultural isolation of Iran in the international community. These are among the achievements of Ahmadinejad.

Iran has a large scientific and academic diaspora (see *Nature* doi:10.1038/news.2009.597; 2009) — how best can it contribute during the current crisis?
Iranian scientists abroad have made, and will continue to make, a constructive contribution by throwing light on the false, suppressive and anti-development policies of Ahmadinejad and his government. The huge scientific capacity of Iranians abroad could also serve well in any serious effort to develop Iran, should a foresighted and compassionate government be established.

In what ways could the international scientific community help colleagues in Iran without meddling — or being seen to meddle — in Iran's internal affairs?
Prominent universities and scientists worldwide can help through legitimate criticism of the Iranian government's treatment of students, academics, scientists and the people of Iran; its violations of citizens' constitutional rights and academic freedom.

I would like to say to my scientific colleagues wherever they are that while Iran, an ancient civilization, has its own cultural background and national interests, the nation and its academics wish to be productive and constructive members of the international community. They reject government adventurism and the creation of a climate of tension internationally.



N. TAVAKOLIAN/POLARIS/EYEVINE

In our last interview, you said that support by Iranian young people and women for the principles of a civil and democratic society were what would "shape Iran's future". Has that shaping moment come?
Yes. The recent growth in Iranian women's and young people's political and social awareness has set the stage for their current demands for greater civil liberties, and structural and democratic reforms. The outcome of the 2005 presidential election, combined with the subsequent mismanagement of the government, has catalysed this process — and explains the 85% voter turnout in last month's presidential election, and the demand for change. The youth, and in particular the supporters of the reformist candidate Mir Hossein Mousavi, played a major role in creating the extraordinary enthusiasm and motivation surrounding the elections. I remain optimistic as to the role, and the movement, of Iran's youth.

How might Iran's legitimate aspirations for nuclear power and technology be reconciled with international concern over its nuclear intentions?
Iran's nuclear issue has become both a political problem and an issue used for domestic political ends, resulting in an artificially tense atmosphere at the international level. But a government that derives its authority from a democratic election is better placed to also build trust at the international level — with a view to exploiting nuclear energy legally and peacefully, to both support the national interest and strengthen international cooperation.

In the 2005 election, you were at one point considered a front-runner, but were beaten in the first round. Were you concerned about the election's legitimacy? How do you plan to pursue your vision of a modern Iran?
I formally protested the illegitimacy and distortion of the first round and called for a ballot recount, but unfortunately the Council of Guardians rejected this request. Distrust towards the outcome of the recent elections has, I think, its roots in doubts with respect to the earlier election. My main mottos in the 2005 election were for the construction of a modern Iran through democracy, scientific development of the country and peace in the world. Fulfilment of the same remains among my basic ideals.

Is an immediate national reconciliation the best way forwards?
There can be no grounds for a national reconciliation in an atmosphere of public distrust. Trust and transparency must first be established by proper examination of the complaints of the presidential candidates. In any case, it seems inevitable that Iran must move towards a political system based on democracy, justice and moderation. ■

Interview by Declan Butler

Additional questions and answers are available at http://tinyurl.com/kupapj.

When Earth greened over

A thick, green carpet of photosynthetic life, on the scale of that seen today, exploded across Earth 850 million years ago — much earlier than thought — a new study suggests.

The matting — a mixture of algae, mosses and fungi — would have fixed atmospheric carbon into the soil, which would then have washed into the seas for burial, according to the study (L. P. Knauth & M. J. Kennedy *Nature* doi:10.1038/nature08213; 2009). With lower levels of carbon to react with, global levels of oxygen would have risen. The greening of ancient Earth could thus be indirectly responsible for the sudden evolution, beginning about 600 million years ago, of larger respirating animals with oxygen-hungry cells, say geologists Paul Knauth of Arizona State University in Tempe and Martin Kennedy of the University of California, Riverside.

"This is a profound event," says Kennedy. "It explains the rise of oxygen, and the timing of that rise."

The evidence the researchers provide is indirect: data compiled from thousands of samples of carbonate rock, such as limestone, that originally formed in ancient shallow seas. Analysis of carbon and oxygen isotopes in these rocks revealed that the influence of the freshwater run-off into these seas was as important in ancient times as it is today in forming carbonate rocks.

Because terrestrial plant life leaves indelible isotopic marks in modern carbonate rocks, the authors surmise that some sort of photosynthetic life — at the same global scale — was responsible for similar measurements they found in ancient rocks. In rocks older than 850 million years, they find starkly different isotopic signatures, which they interpret as an absence of carbon-rich material in freshwater run-off, and thus an absence of photosynthetic life on land.

H.STRAND/CORBIS

Did a green carpet of algae, mosses and fungi cover Earth 850 million years ago?

The study contradicts other work that looks to the oceans, rather than land, to justify the same isotopic data. Other researchers argue that the oxygenation of Earth and the explosion of animals 600 million years ago arose from sudden and drastic changes in ocean water chemistry around the same time (D. A. Fike *et al.* *Nature* 444, 744–747; 2006). The changes in ocean chemistry have been attributed to episodic releases of methane from ocean vents, and periods of 'snowball Earth', extreme glacial epochs when Earth may have been so cold that oceans froze over.

But Knauth and Kennedy say the isotopic records in the carbonate rocks reflect more than just the chemistry of the global oceans. They argue that most carbonate rocks undergo further stages of alteration where freshwater run-off is important.

There are problems with the new theory, says Paul Falkowski, a geochemist at Rutgers University in New Brunswick, New Jersey — most notably that there isn't much evidence for widespread plant life until around 400 million years ago. The hard tissues of vascular plants evolved around this time, but the softer tissues of mosses and fungi that came before would have been preserved less easily. Work with molecular clocks — which use genetic differences to estimate the timing of speciation — does suggest that terrestrial plants evolved from the types of plants that Knauth and Kennedy call for, around the time that they suggest.

But to have the effect on the carbonate record that they see, the ancient photosynthetic life would have needed to be operating on the scale that it is today — a worldwide carpeting of green. And that should have left something for posterity, says Nick Butterfield, a palaeobiologist at the University of Cambridge, UK. "In order to have a significant impact it has to be everywhere, all over the place," he says. "And it can't be, unless it has seeds and cuticles and adaptations for covering vast amounts of the terrestrial surfaces. If you've got those adaptations you can't avoid turning up in the fossil record." ■

Eric Hand

US AIDS chief lays out priorities

The new head of the US President's Emergency Plan for AIDS Relief (PEPFAR) takes the reins of the sometimes controversial programme just as the global economic slump has made the job more difficult than ever.

The Senate confirmed physician Eric Goosby as US global AIDS coordinator on 19 June. Goosby was deputy director of the White House Office of National AIDS Policy during the Clinton administration, and then chief executive officer of the non-profit Pangaea Global AIDS Foundation in San Francisco, California.

Doctors and advocates hope that Goosby will expand PEPFAR's success in delivering AIDS treatment and services while ditching less favoured parts of the programme, such as its focus on prevention by the ABC strategy — abstain, be faithful, use condoms.

"He is incredibly well-grounded in the prevention and treatment issues facing the developing world," says Christine Lubinski, vice-president for global health at the Infectious Diseases Society of America in Arlington, Virginia. "But he has got an enormous challenge ahead of him."

In May, advocates called President Barack Obama's global health request of $63 billion over six years "meager", saying it broke his campaign pledge to provide $1 billion a year in new spending for PEPFAR. The plan spent $18.8 billion between 2003 and 2008, making it the largest aid programme for one disease.

> "Goosby has got an enormous challenge ahead of him."

The recession has forced some aid recipients to scale back. In February, Botswana — which has received more than $300 million from PEPFAR — said that after 2016 it might stop enrolling patients in its programme to provide free medications. And in May, Tanzania, a recipient of $828 million from PEPFAR, said it would shrink its AIDS programmes by 25%.

"The sustainability issue is of high concern for us," Goosby acknowledges. He hopes to shift PEPFAR's focus away from US non-governmental organizations and towards helping aid-recipient governments set their priorities.

His other main priority is prevention. PEPFAR has been criticized for focusing on ABC because it fails to help many groups at risk in developing nations, such as married women. Goosby says PEPFAR's prevention efforts must be more sustained and aggressive, should target high-risk groups and should broaden their focus — by, for instance, making condoms and reproductive-health services more widely available.

These priorities are encouraging to doctors and advocates such as Paul Volberding, vice-chair of medicine at the University of California, San Francisco. "I think that there will be some big changes," he says. ■

Erika Check Hayden

Evolution wins out in Hong Kong curriculum dispute

Hong Kong's Education Bureau has vindicated biologists — and disappointed creationists — by clarifying its guidelines for teaching biology in secondary schools.

Debate had been raging since February over the revamped curriculum, to go into effect in September, which "in addition to Darwin's theory" encourages the teaching of "other explanations for evolution and the origins of life". Some scientists worried the clause implied support for creationism and intelligent design, which are already taught in many of Hong Kong's religious schools (see *Nature* 457, 1067; 2009).

A five-page clarification, released on 22 June, said that studies of the origin of life should focus on Darwinism and that "other explanations" referred to alternative theories of evolution, such as those put forth by Jean-Baptiste Lamarck and Alfred Russel Wallace.

The biology curriculum does not include the teaching of creationism, intelligent design and "non-scientific explanations", the document says. Teachers may still choose to cover topics beyond the guidelines, however.

Japanese diplomat chosen to lead nuclear watchdog

The International Atomic Energy Agency (IAEA) voted in Japan's Yukiya Amano as its next director general on 2 July. Amano, 62, won the necessary two-thirds majority of the IAEA's 35-member board by a single vote to beat chief rival Abdul Samad Minty of South Africa, considered a favourite of the non-nuclear-weapons countries.

Based in Vienna, the IAEA is the international body charged with policing the nuclear-nonproliferation treaty (NPT), the primary agreement meant to halt the spread of nuclear weapons.

Amano is a lawyer by training and has an extensive background in disarmament and nonproliferation. Pending final approval in September by the 146-member IAEA general conference, he will succeed Mohamed ElBaradei, who has led the agency since 1997 and who along with the agency received the 2005 Nobel Peace Prize. With a reputation for a quiet, low-key style, Amano is unlikely to have the profile of ElBaradei, whose occasionally inflammatory comments made headlines.

For a longer version of this story, see http://tinyurl.com/ndz8o8

D. CALMA/IAEA

Yukiya Amano will head efforts to police the spread of nuclear weapons.

Lunar Reconnaissance Orbiter snaps test pictures

All is going smoothly so far for NASA's Lunar Reconnaissance Orbiter, which launched on 18 June. On 2 July it captured its first high-resolution test images of the Moon (pictured right), kicking off a year-long mapping mission.

Regular updates from the orbiter can be found on its Twitter feed (http://twitter.com/LRO_NASA). Meanwhile, *Nature* is twittering the *Apollo 11* Moon mission as it happened in real time — 40 years on (http://twitter.com/ApolloPlus40).

NASA/GSFC/ARIZONA STATE UNIV.

Governments fail to reduce global biodiversity decline

Governments worldwide are failing their commitments to reduce biodiversity loss by 2010, according to a report released last week by the International Union for Conservation of Nature (IUCN).

The IUCN's analysis of its 2008 global 'Red List' of threatened species uses indices to track trends in extinction risk. Birds, mammals, amphibians and corals all showed a continuing deterioration, it says.

The United Nations has declared 2010 the international year of biodiversity, under its Convention on Biological Diversity. On 26 June, European Union (EU) environment ministers meeting in Luxembourg said they were "deeply concerned" by a European Commission assessment in December stating that the EU is unlikely to meet its own ambitious aim of halting biodiversity loss by 2010.

US scientist jailed for sharing sensitive data

A former University of Tennessee professor has been sentenced to four years in prison for sharing sensitive technologies with his Chinese and Iranian graduate students.

J. Reece Roth, an emeritus professor of electrical engineering, was sentenced on 1 July by a Tennessee district court for violating the Arms Export Control Act. He had been developing ways to reduce the drag on unmanned planes, and employed two research assistants without obtaining the required licence (see *Nature* 442, 232–233; 2006). Roth plans to appeal the verdict.

In a separate case, a Chinese-born scientist who has lived in the United States for 23 years is suing the US government for rights violations for expelling him last year from the NASA Ames Research Center, California.

Haiping Su, a US citizen who received his doctorate in 1991 from Kansas State University in Manhattan, alleged in a case filed on 24 June in a San Jose federal court that a 2007 security badge-issuing process led to his illegal ousting.

Su was working on airborne systems for imaging forests. His attorneys say he had no involvement with classified material.

US Air Force will continue to share meteor data

The United States Air Force says that it will resume sharing data on incoming meteors with astronomers.

The Air Force collects the data with a network of satellites and sensors designed as a missile early-warning system. For more than a decade, it provided them to astronomers on an ad-hoc basis, but the informal relationship came to a halt earlier this year (see *Nature* 459, 896–897; 2009).

Astronomers feared that the Air Force had put a stop to the practice, but "the data will still flow", says Andy Roake, a spokesman for Air Force Space Command in Colorado Springs, Colorado. The Air Force is developing procedures for releasing data that will be faster, more systematic and in compliance with classification procedures. Data sharing could resume within the next few months.

NATURE|Vol 460|9 July 2009

ONE GENE, TWENTY YEARS

When the cystic fibrosis gene was found in 1989, therapy seemed around the corner. Two decades on, biologists still have a long way to go, finds **Helen Pearson**.

During the day, Lap-Chee Tsui and Francis Collins were attending a gene-mapping workshop. At night they were scrutinizing the pages churning out of a fax machine they had set up in a dorm room. Their hunt for the cause of cystic fibrosis had reached a gene that looked from its sequence like it might have a role in transporting ions through cell membranes, a process that goes awry in those with the disease. The fax they received that night from Tsui's lab showed that many people who have cystic fibrosis lack three base pairs from both copies of this gene, whereas those without the disease always have at least one copy intact. With that fax, on a rainy night in May 1989, "I was convinced — that was the moment," Collins says.

Four months later a four-year-old boy with cystic fibrosis, Danny Bessette, was shown sitting cross-legged on the cover of *Science*, framed by a rainbow of chromosomes. Inside the magazine, three papers[1–3] laid out the details of the discovery of the gene responsible for Bessette's condition — the first gene for a human disease discovered without the help of an already-known protein sequence or any clue to its whereabouts. "In this issue ... there is a story that does not begin at the beginning or end at the end, but has a very happy middle," wrote *Science*'s editor Daniel Koshland[4]. "One in 2000 children born each year with a fatal defect now has a greater chance for a happy future." By that stage, news of the finding had already leaked to the media, been the subject of two hastily assembled press conferences and been trumpeted in newspapers worldwide. "It would be difficult to overstate the importance of the cloning of the cystic fibrosis gene," wrote geneticist Peter Goodfellow in *Nature* that month[5]. "The implications of this research are profound: there will be large spin offs in basic biology, especially in cell physiology, but the largest impact will be medical."

So far, Goodfellow's prediction has proved wrong, at least as far as medical impact is concerned. As Jack Riordan, who collaborated with Tsui and Collins on the original discovery, puts it: "The disease has contributed much more to science than science has contributed to the disease."

This is not to deny that medical progress has been impressive. An American born with cystic fibrosis today has a life expectancy at least ten years longer than one born in 1989 did. Such advancements help explain why Bessette — now 24, and pictured opposite — has a future at all. But many researchers concede that relatively little of that improvement can be laid at the door of the cystic-fibrosis transmembrane regulator gene, or *CFTR*. Gene therapy — the source of so much of the hope in 1989 — has so far bought no one with this condition a single additional year of life; no therapies targeted at the CFTR protein have yet been approved. Researchers have not even fully agreed on a hypothesis to explain how mutations in the gene cause the condition. But the gene itself "found its way into all departments", says Riordan, leading to progress in fields as diverse as protein trafficking and membrane transport. And the gene-hunting techniques that Tsui, Collins, Riordan and their colleagues pioneered have laid the foundation for a genetic understanding of all human disease.

"The disease has contributed much more to science than science has contributed to the disease."
— Jack Riordan

Twenty years, although a long time in the life of a young man such as Bessette, is not the whole story. Several hundred million dollars have been spent trying to find a therapy that directly tackles the molecular defects that underlie cystic fibrosis; Collins, for one, thinks that this means the hopes on which gene therapy never delivered are about to be fulfilled. Like many researchers, he is excited by clinical results coming through on a pair of small molecules that could get mutant versions of the CFTR protein to work properly. Should the molecules be approved, "it will be a pair of home runs, a milestone for all genetic disease", Collins says. And those home runs would never have been hit without the gene and the opportunity to study the protein that needs fixing. "You can paint a direct pathway from the gene discovery [to those drugs]," he says.

To call the path direct might be overstating it. Researchers have taken many paths from *CFTR*, and their travels have shown that behind this gene and every one found since lie dauntingly complex biological stories. "I think one of the lessons of cystic fibrosis is the recognition of the enormous challenge that faces us in human biology," says Riordan, now at the University of North Carolina, Chapel Hill. "It's not like going to the Moon — it's going to Mars." The size of the challenge can sap enthusiasm. "Looking back, it was an important contribution," says Tsui, "but I'm disappointed because at this time, from my own research, I was not able to help very much." Riordan says that he now views "the latest hot gene" with a "jaundiced eye". But one thing that shines through when speaking to these three and other researchers is their continued optimism, their passion and their sense of urgency. "Perhaps," says Collins, "we've taken our blinkers off. Perhaps we couldn't deal with it before, and now we have a lot more tools to dissect the complexity."

"It's not that it hasn't worked," says Riordan. "It's only been 20 years."

Blind beginnings

Geneticists have been interested in cystic fibrosis since the disease was first identified in the 1930s. The disease is common in Caucasian populations — about 1 in every 25 people carries a mutated copy — and its pattern of inheritance is straightforwardly Mendelian: those with one mutated gene are healthy carriers; those who inherit two will have the condition. Doctors knew that although the pancreas often fails and the gut is unable to absorb nutrients, the lung is the organ that is crippled with recurrent and persistent infections, "and that's unfortunately the one that kills them", says Richard Boucher, a pulmonary physician and cystic fibrosis researcher at the University of North Carolina. But for

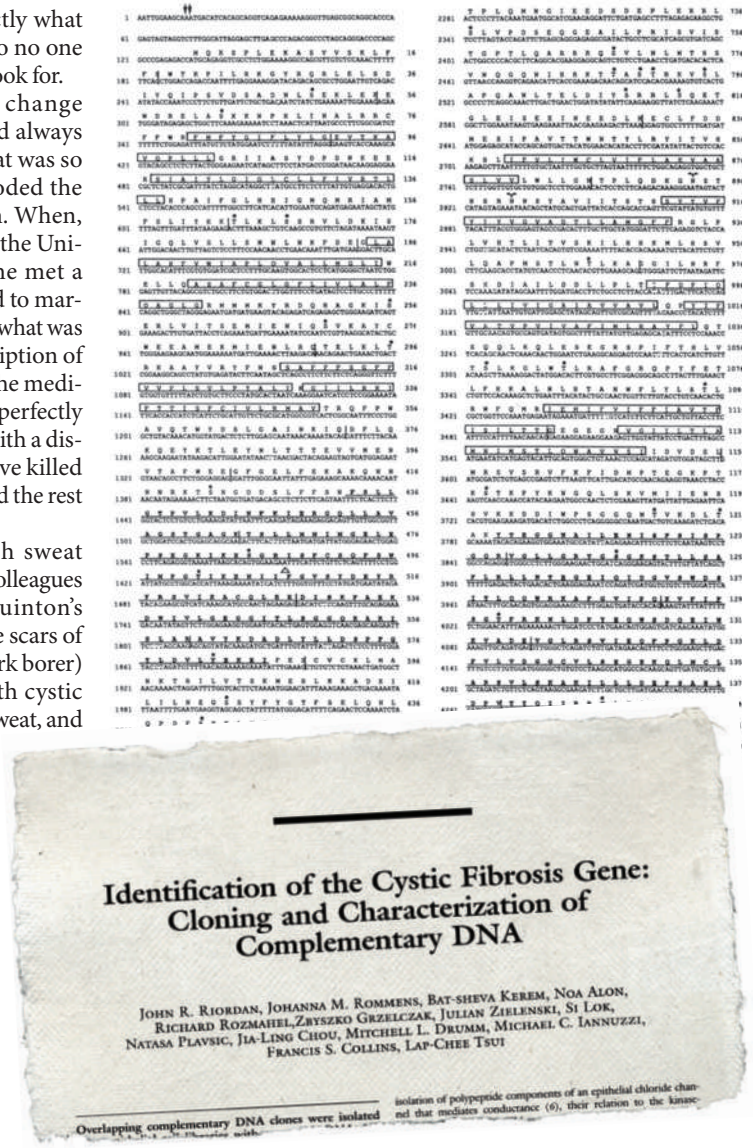
J. LUIS MAGANA/AP

decades no one knew exactly what was wrong with the cells, so no one knew what type of gene to look for.

Paul Quinton helped change that. As a kid, Quinton had always coughed a lot, and his sweat was so salty that his clothes corroded the wire hangers they dried on. When, in 1965, as a 19-year-old at the University of Austin, Texas, he met a girl and his thoughts turned to marriage, he decided to find out what was wrong with him. The description of cystic fibrosis he found in the medical library fit his symptoms perfectly and he diagnosed himself with a disease that should already have killed him, but that he would spend the rest of his life studying.

Quinton collected fresh sweat glands from visitors, from colleagues (Riordan, who visited Quinton's ranch, says he still bears the scars of Quinton's biopsies with a cork borer) and from other people with cystic fibrosis to explore why his sweat, and that of others with the disease, was so salty. In 1982, while working at the University of California, Riverside, an experiment measuring the ability of sodium and chloride to pass through the glands led him to finger a channel that was unable to conduct chloride ions across the epithelium of the skin, and that might also underlie problems in the lungs and the other affected organs[6]. "I feel silly saying it but I literally jumped up and ran up and down the hall shouting 'Eureka'," says Quinton, who now also works at the University of California, San Diego. "I still get chills; it was one of those moments you get once in a lifetime." The disease is evident when he speaks: he still clears his throat and coughs a lot.

Quinton's discovery and others like it told geneticists what they should be looking for: a gene that is involved in the movement of chloride, and perhaps other ions, across the epithelium. By now an intense and competitive hunt was under way. It was the 1980s, when the human genetic sequence was largely uncharted territory, and the human genome project was still a twinkle in various eyes, including Collins's. Finding the gene would be a technical and intellectual challenge as well as a medical breakthrough. Until that point, almost all of the genes that had been associated with human diseases had been identified by first isolating the protein responsible. A protein's amino-acid sequence reveals much of the gene's probable nucleotide sequence, and that made pinpointing the gene easier. The few exceptions, such as those found for Duchenne muscular dystrophy and retinoblastoma, were helped by a few patients with chromosomal abnormalities that pointed to the gene's position. For cystic fibrosis, researchers were working blind: they had no protein and no location. This was to be a big test of new 'reverse genetics' techniques, in which a gene is found by searching for markers in the genome that are consistently inherited with the disease in affected families and using them as signposts to the gene itself.



Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA

John R. Riordan, Johanna M. Rommens, Bat-sheva Kerem, Noa Alon, Richard Rozmahel, Zbyszko Grzelczak, Julian Zielenski, Si Lok, Natasa Plavsic, Jia-Ling Chou, Mitchell L. Drumm, Michael C. Iannuzzi, Francis S. Collins, Lap-Chee Tsui

Overlapping complementary DNA clones were isolated ... isolation of polypeptide components of an epithelial chloride channel that mediates conductance (6), their relation to the kinase-

The sequence of *CFTR* revealed the position at which three base pairs are commonly deleted (triangle).

Tsui, then at the Hospital for Sick Children in Toronto, Canada, and now at the University of Hong Kong, was a key player in the hunt; so were Robert Williamson at St Mary's Hospital Medical School in London, and a handful of other researchers. By 1985, several groups[7–9] had shown that the gene mapped to a region of chromosome seven, but it was still a vast genetic wilderness somewhere between one and two million base pairs wide. In 1987, Williamson announced that he had landed on the gene, but soon after had to admit he had got it wrong. Nevertheless, many groups assumed that Williamson was close and dropped out of the race at that point. Says Collins: "Lap-Chee and I were more stubborn".

Collins, then at the University of Michigan in Ann Arbor and until last year the head of the National Human Genome Research Institute in Bethesda, Maryland, met Tsui at that year's meeeting of the American Society of Human Genetics. A few years previously, Collins had described the technique of 'chromosome jumping', a way of leaping across the vast genetic distances from one marker sequence in a region to another that was much faster than the conventional way of chromosome 'walking'[10,11]. They agreed to collaborate: Collins's lab would bound to new positions, and Tsui's would walk forwards and backwards from the landing points looking for the gene. Two years later, on that rainy night in the dorm room, their fax machine told them they had found it.

The *Science* papers showed that the gene looked like others encoding membrane proteins that transport ions. The three base pairs missing in the vast majority of people with cystic fibrosis eliminated an amino acid at position 508 of the protein's amino-acid sequence, a mutation called ΔF508. "It was exciting times," says Robert Beall, then executive vice-president for medical affairs at the Cystic Fibrosis Foundation in Bethesda, Maryland, and now its director. "We had been at a bottleneck. We didn't know why chloride wasn't getting out of cells, and that gene solved it."

The scramble of competition continued as researchers

rushed to work with the gene. John Hanrahan at McGill University in Montreal, Quebec, recalls the time he was collaborating with Riordan on a paper for *Cell*[12]. Riordan called him to ask him to fax through a figure for the manuscript as he was worried about a scoop from a competing paper at *Science*. "I raced to the airport in a snowstorm to send the originals by same-day courier, but it was the faxed version that went to press," Hanrahan says. "When people look at the traces they must wonder, 'Why are they so pixelated?'" But Hanrahan, like most researchers, says that the competition was a healthy one, even if it deprived them of a little sleep. "I think a lot of data were published and some mistakes were made, but there was tremendous excitement and the field moved ahead rapidly."

From the beginning, the goal was gene therapy. Get a good gene into the patients and they would make the proper protein; with the proper protein they'd be cured. But the path from gene to therapy wasn't smooth. It took more than a year just to get bacteria to produce the protein from the cloned gene, because of 'cryptic' sequences within the gene that prevented the bacteria from expressing it. But by 1993 the first clinical trials were under way.

"The expectation was that all you needed to do was get a little bit of stuff to act in the lungs and 'hey presto' you'd have a Nobel prize," says Steven Hyde, who works on cystic fibrosis gene therapy at the University of Oxford, UK. Among other things, the lung, researchers now realize, is just about the worst possible target for such an approach. Its sophisticated defences against infection have evolved precisely to prevent the sort of uptake and expression of foreign material the gene therapists were after. Mike Welsh at the University of Iowa and his colleagues, who in one of the first trials pushed the gene into cells in the nasal passage as a surrogate for those in the lung, later realized that the cells that had taken up the gene were probably damaged during the procedure. "A whole slew of people did similar trials and everyone got a little disillusioned." Hyde says.

Disillusionment isn't enough to kill off an idea — but death is. In 1999, a severe immunological reaction killed Jesse Gelsinger in a gene-therapy trial for an inherited liver disease, casting a pall over the entire field. In the United States, the field has never fully recovered. In other countries — the United Kingdom and France, for example — researchers have been much more active in pursuing the technique.

Around the same time, Beall decided to turn the gene into a way to find a therapy, rather than being the therapy itself. He wanted to take advantage of new tools coming online for high-throughput drug screening. Researchers inserted the gene into cells, expressed the mutated protein, then screened for drugs that could correct the way the protein is made or the way it works. "People thought we were crazy," Beall says. What started as US$2-million grant in 1999 has turned into an $76-million programme, and Beall proudly points to a chart showing the drugs working their way through the pipeline as a result.

By far the most common mutation is ΔF508. It causes the protein to fold up poorly, and a drug known as a corrector is needed to help it fold correctly and get to the membrane it needs to sit in. Other mutations — there are now more than 1,500 known in the gene — require different approaches. Versions of the gene in which protein translation stops short need drugs to override the stop signal. Then there are proteins that get made, fold up and reach the membrane but just don't work properly. They need what are called potentiators.

Lap-Chee Tsui, Francis Collins and Jack Riordan (left to right) celebrate their 1989 discovery of the cystic fibrosis gene with a patient.

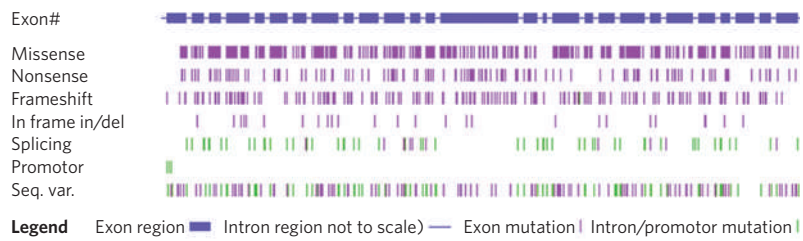
CANADIAN CYSTIC FIBROSIS FOUNDATION

"Perhaps we couldn't deal with it before, and now we have a lot more tools to dissect the complexity."
— Francis Collins

In March 2008, investigators presented results from a phase II trial of the potentiator VX-770 to a room of several hundred researchers at a meeting of the Cystic Fibrosis Foundation. Just two weeks of treatment in 20 people with a rare mutation called G551D had dramatically lowered some people's sweat chloride and produced some improvement in lung function — something that clinicians found particularly remarkable given the battered state of their airways. "When they showed those data and I saw the emotions from those physicians, it was unbelievable," Beall says. "It was the most emotional time since the discovery of that gene. It's telling you we can change the course of this disease." Collins agrees. "It was wildly better than even the most optimistic perspective for a small-molecule trial," he says. Phase III trials of VX-770, developed by Vertex Pharmaceuticals of Cambridge, Massachusetts, are now recruiting patients. Beall and others say that the drug might find a much wider market if it is also used in people with other mutations, including ΔF508, in conjunction with a corrector. That corrector could be another Vertex drug called VX-809, which is just starting phase II trials.

A special case

Must it take 20 years to get from gene to drug? No. Various things have made cystic fibrosis peculiarly difficult. One has been a lack of a complete understanding of how the CFTR protein leads to the disease. Many think that the defective channel causes the lungs to absorb too much water; others have argued that the primary problem is an incorrect ion



More than 1,500 mutations of various types have been found in the *CFTR* gene.

composition that disables the lungs' normal defences against infection. This debate became so fierce it was described as the 'salt wars'. At least part of the problem seems to lie in another ion channel that CFTR interacts with. "If you ask 20 people you'll get 20 different hypotheses," says Welsh. "Everybody's got their favourite — I think we don't know."

Then there are some purely technical problems. Mice with mutated versions of *CFTR* have few obvious lung problems and thus make poor models of the disease. (The models have, however, revealed something about why the mutated gene is so common — see 'A killer advantage'.) The CFTR protein is huge and is embedded in a membrane, making its structure difficult to determine with X-ray crystallography; plus the fact that airway cells tend to contain only a hundred or so copies of the protein, so there is very little of the stuff to play with. Together, these mean that no one has been able to resolve a complete high-resolution structure for the protein, which has hampered understanding of how it works and the design of drugs.

Other genes have had it easier. Collins points to the gene for Hutchinson–Gilford Progeria Syndrome (HGPS), an extremely rare single-gene disease that causes young children to shows signs of old age. The gene was discovered by Collins's team at the National Human Genome Research Institute in 2003 (ref. 13) and by another group in France[14], and a treatment based on it went into a phase II clinical trial in 2007 — a notably fast pace of translation. Collins puts much of the speed down to serendipity. The mutated protein was an extremely well-studied one called lamin A, and a cancer drug that had already reached late-stage clinical trials was found to work against the mutated protein, saving some laborious drug screening and safety testing. What's more, the task required of the drug is simpler. Drugs for cystic fibrosis have to compensate for or restore the function of a mutated protein, whereas those for HGPS simply have to block the action of one that has turned toxic.

"If you ask 20 people you'll get 20 different hypotheses. Everybody's got their favourite."
— Mike Welsh

The discovery of *CFTR* deserves at least some credit in the HGPS story, though, as it does for accelerating the pace of translation after almost every gene discovery since 1989. That's because hard work and mistakes made in this field have saved effort in every other. "If you found a new gene tomorrow you could compress those 20 years hugely because of what's been done with cystic fibrosis," says Hyde.

And gene therapy may yet prove possible for cystic fibrosis. In 2001, the Cystic Fibrosis Trust in Bromley, UK, asked Hyde's group and two others in Britain that were still working in the field to stop competing and start working together. They complied and have spent several years and around £30 million (US$49 million) working methodically through some of the problems — such as devising better ways to measure changes in lung function. Earlier this year, researchers at Imperial College London treated the first of 27 people with cystic fibrosis in what is expected to become the largest gene-therapy trial ever undertaken for the disease. The aim is to test whether the gene can be delivered safely, in a fatty particle called a liposome. If it is, the researchers will scale up to a 100-person randomized controlled trial to see whether it is effective. "I think we have now tempered the optimism of the early 90s with a heavy dose of realism," says Eric Alton, who directs the trial.

Clinical changes

Throughout this time there have been dramatic changes in the way that cystic fibrosis is treated in the clinic. In 1994, Genentech introduced Pulmozyme (dornase alfa), an enzyme that breaks up some of the lung-clogging mucus that encourages infections. A few years later, aerosolized antibiotics were introduced to fight these infections more aggressively. Earlier this decade doctors in Australia started noticing that their patients who surfed felt better during the surf season — leading researchers to test the idea that the daily inhalation of super-salty water, called hypertonic saline, could help lubricate the lungs. It did[15,16], and this is now standard therapy for many patients. Not all of them benefit from this approach, though: Bessette stopped taking hypertonic saline after a few years because it made him cough blood from a burst vessel in his lung. Pulmozyme does

A killer advantage

Work with the cystic-fibrosis transmembrane regulator gene, or *CFTR*, has helped to solve one puzzle about the disease: why do so many people carry a mutated version? For most of human history, people who inherited two mutated copies died long before they reached reproductive age — and if they didn't, men with cystic fibrosis tend to be infertile because they lack the sperm-carrying vas deferens. This means that selection should long ago have weeded out the mutations. The fact that it has not has led to the hypothesis that the mutated gene has a 'heterozygous advantage': people who carry one copy (heterozygotes) have some selective advantage that keeps the frequency of mutated alleles high.

One idea, that heterozygotes might enjoy some kind of increased fertility, has prompted numerous studies but most of them have been inconclusive. Then, in 1988, geneticist Lynn Jorde from the University of Utah in Salt Lake City more or less killed the hypothesis with a more rigorous investigation, showing in a Mormon population with good family histories that cystic-fibrosis carriers bore no more children than those without the mutation[18].

Jorde says that the best evidence for a heterozygous advantage has come from one of the mouse models. The bacterium that causes typhoid fever uses *CFTR* to enter cells, and mice heterozygous for the mutated gene seem to be resistant to infection[19]. If the same is true in humans, then heterozygotes would have had a selective advantage during typhoid epidemics. "Typhoid has been around a long time and killed a lot of people. If humans had resistance it would be a strong selection factor," Jorde says. That selection pressure would no longer be so prevalent, however, as the disease is controlled by vaccination in many countries. **H.P.**

feature in his 40–50 pill-per-day regime, and he anticipates more advances that might improve his lung function. "Yes, we all hope for a cure, but if they can just help us stay healthy that in itself is quite an accomplishment," Bessette says.

Quinton, too, follows a rigorous regimen, inhaling hypertonic saline every day and taking intravenous antibiotics every few months. He rides his bike to work, but he can't run far or play basketball. Both upper lobes of his lung have been removed because of chronic inflammation. For someone born when he was, though, things could have been much worse — and thanks to research into the *CFTR* gene, Quinton knows why they're not. Although he has one copy of ΔF508, the mutation in his other gene, R17H, has relatively mild effects. He found this out when, several years after the gene was found, Garry Cutting at Johns Hopkins University School of Medicine in Baltimore, Maryland, analysed his genes as part of work on genetic testing for the disease.

These tests are "probably the most common form of genetic testing in the world today", says Cutting, who now drafts clinical-testing guidelines for cystic fibrosis. In the United States and some European countries many pregnant women and their partners are offered testing for mutations in *CFTR*, forcing clinical geneticists to confront issues about genetic counselling and genetic risk that are likely to escalate as more and more genes become as well studied. Working with a gene that has so many mutations, most of which are still little understood, underlines the futility of testing for something with no known clinical severity and therefore no rational basis on which to make decisions about ending a pregnancy. "The agony I've seen for some couples where one is a carrier and one has a mutation of unknown significance," Cutting says, "it is just immense." Newborn screening, which is also commonplace in some countries and typically involves a biochemical test followed by a genetic one, throws up similar issues for clinicians who may be unable to advise parents how severely their child is likely to be affected.

New studies are making the molecular landscape look even more complicated. Two years ago, the Cystic Fibrosis Foundation helped to launch a North American consortium to search for 'modifier genes' at work in the disease that might explain why some people with two copies of the ΔF508 mutation die at 16 whereas others have pretty healthy lungs into their 20s. The consortium members recently screened more than 4,500 people to look for genetic variations that are strongly linked with severity of the disease, says consortium member Michael Knowles from the University of North Carolina. One of the strongest variants to have emerged from previous studies of modifier genes, called TCF7L2, is also thought to strongly predispose carriers to type 2 diabetes[17]. The link may lie in the failure of the pancreas and consequent diabetes that cystic fibrosis frequently causes.

Results such as these suggest that once the *CFTR* gene and its protein are viewed in context, cystic fibrosis will spiral into a new realm of dizzying complexity. If studies of one gene have expanded to fill 20 years, how many years can be filled once the tens or even hundreds of modifier genes are factored in, let alone whatever other influences there may be outside the genetic code? For Knowles, though, the results present an exciting opportunity rather than a daunting complexity. He sees cystic fibrosis as "leading the way" for researchers investigating more genetically complex diseases. If he and others can get to grips with the numerous mutations in *CFTR* and its modifiers, they say that cystic fibrosis could serve as a case study for personalized medicine. Newborns identified with the disease could have their *CFTR* gene and other major modifier genes analysed to choose the most appropriate therapies — assuming, that is, that such a range exists by that point.

R. G. CRYSTAL/NIH

Early gene-therapy trials showed how difficult it is to express foreign genes in the lung.

"We have now tempered the optimism of the early '90s with a heavy dose of realism."
— Eric Alton

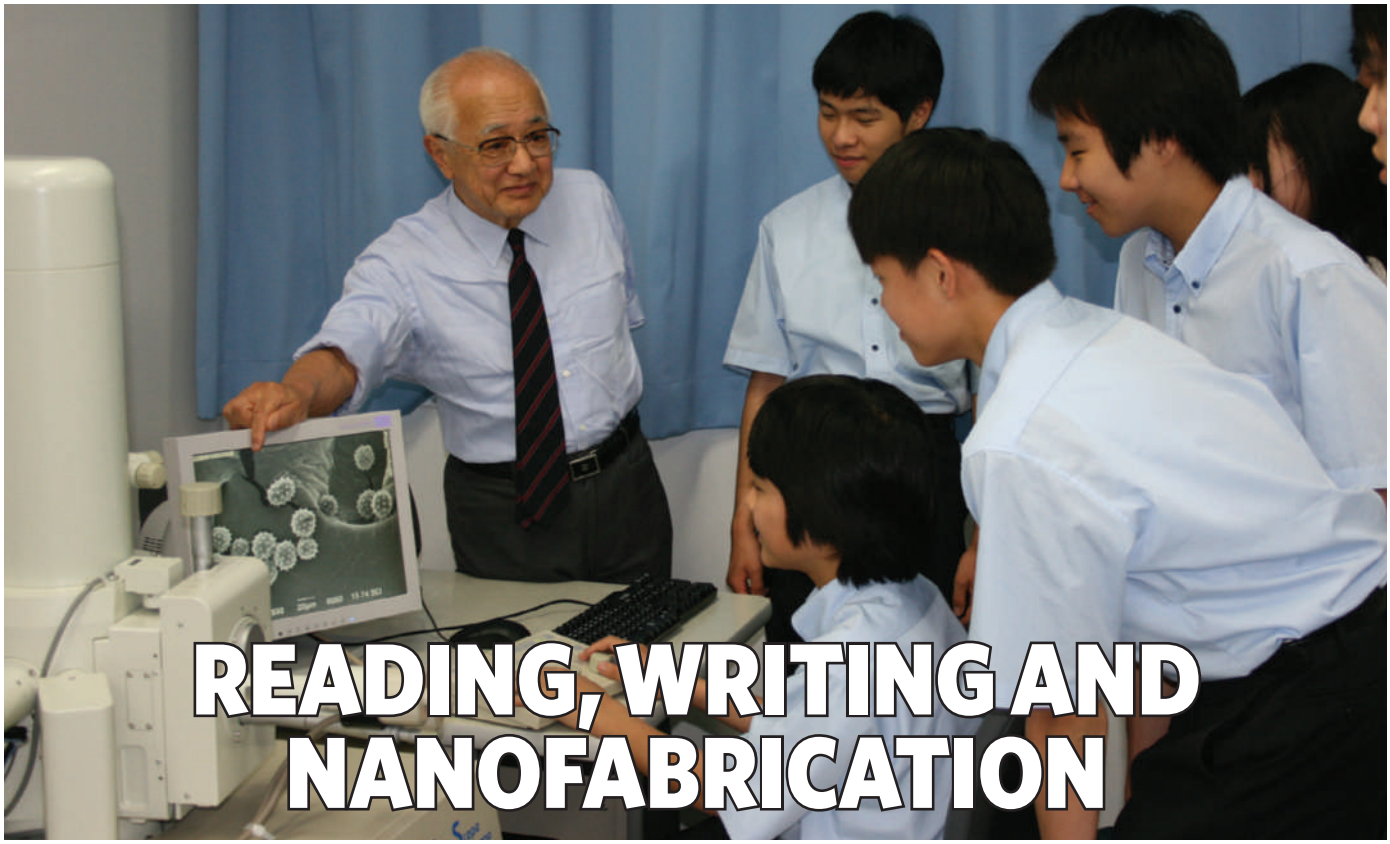
Although he has discovered molecular truths about himself that he might never have expected, "it would be hard for me to say I have benefited from the work I've done", Quinton says. Nonetheless, he, Riordan and others whose careers in this field stretch back farther than 1989 are still hopeful. Like most researchers and clinicians, they are focused on what they can achieve in the next 2–5 years, not what they should have achieved already. "I'd say don't give up," Quinton says. "This really is the only solution. As we succeed on one platform, it will make it much easier to succeed on another."

"It's a helluva lot more complicated than we realized," he says. "We went to the Moon in '69 and the conceit was we could do anything — we corrected polio, we wiped out smallpox. But when you start taking the system apart we've been really naive."

"But that's biology — it's not fair." ■

Helen Pearson is *Nature*'s chief features editor.

1. Rommens, J. M. *et al. Science* **245,** 1059–1065 (1989).
2. Riordan, J. R. *et al. Science* **245,** 1066–1073 (1985).
3. Kerem, B. *et al. Science* **245,** 1073–1080 (1985).
4. Koshland, D. E. *Science* **245,** 1029 (1989).
5. Goodfellow P. N. *Nature* **341,** 102–103 (1989).
6. Quinton, P. M. *Nature* **301,** 421–422 (1983).
7. Tsui, L.-C., *et al. Science* **230,** 1054–1057 (1985).
8. Knowlton, R. G. *et al. Nature* **318,** 380–382 (1985).
9. Wainwright, B. J. *et al. Nature* **318,** 384–385 (1985).
10. Collins, F. S. & Weissman, S. M. *Proc. Natl Acad. Sci. USA* **81,** 6812–6816 (1984).
11. Collins, F. S. *et al. Science* **235,** 1046–1049 (1987).
12. Kartner, N. *et al. Cell* **64,** 681–691 (1991).
13. Eriksson, M. *et al. Nature* **423,** 293–298(2003).
14. De Sandre-Giovannoli, A. *et al. Science* **300,** 2055 (2003).
15. Elkins, M. R. *et al. N. Engl. J. Med.* **354,** 229–240 (2006).
16. Donaldson, S. H. *N. Engl. J. Med.* **354,** 241–250 (2006).
17. Blackman, S. M. *et al. Diabetologia* (in the press).
18. Jorde, L. B. & Lathrop, G. M. *Am. J. Hum. Genet.* **42,** 808–815 (1988).
19. Pier, G. B. *et al. Nature* **393,** 79–82 (1998).


YSFH

READING, WRITING AND NANOFABRICATION

With its electron microscope, genetic sequencing machines and observatory, the Yokohama Science Frontier High School is equipped like no other. Will future scientists be inspired there, asks **David Cyranoski.**

The timetable for 15-year-old students at Yokohama Science Frontier High School (YSFH) can be busy. Before break, they might grow single-layer carbon nanotubes in argon gas and evaluate them with micro-Raman spectroscopy. After break, there are polymerase chain reactions (PCR) to be done. Things don't slow down after class, when students stick around in the observatory to glimpse star clusters or Saturn's rings from the school observatory. The equipment list at Japan's first dedicated science high school, which began classes in April, could rival a small research institute.

The school is also a sophisticated experiment. Its main champion, biophysicist and genomics pioneer Akiyoshi Wada, hopes a "flood" of such institutes will open up throughout the country, inspiring students and Japan's future leaders. Wada thinks that the school is key to reversing children's waning interest in mathematics and science, a phenomenon that has attracted political hand-wringing and has even been given a name — *rika banare*.

As one of five 'super advisers' to the school, and the only one with a permanent position there, Wada has been instrumental in its creation. He spent most of his four-decade career initiating, managing and administrating ambitious science projects aimed at keeping Japan at the forefront of international scientific trends. A decade ago, when the city of Yokohama asked Wada to be on the planning committee for a new high school, he brought the same bold, uncompromising vision. "It was precisely because of Dr Wada that the school was able to establish its educational principles and goals," says the school's principal, Haruo Sato.

There are wrinkles to be ironed out. The school will be open to accusations of elitism and, with a price tag of ¥9.5 billion (US$100 million) — not including the land, which was donated by the city — some people ask whether the model really has a chance of spreading.

Wada answers yes to the question before it is even finished. But he acknowledges that the Yokohama experiment has much to prove. "The future of science education in Japan will depend greatly on the success of the YSFH," he says, "and I am aware of that massive responsibility."

Ancient roots

Wada is softly-spoken. On a tour of the five-floor, 25,000-square-metre buildings that overlook the Tsurumi River, he trails behind letting an English teacher, Yukimasa Uekusa, lead. Wada seems proudest of the finer details, such as the two famous trees outside — a descendant of the apple tree in Isaac Newton's garden and an offshoot of the grape vine used in some of Gregor Mendel's experiments — and the larger-than-life images of famous scientists that cover many of the walls. "The captions are all in English," he says. "They need to learn English." Besides spending 2–3 days a week at the school, Wada also runs the 'Wada Salon' where he discusses, over tea and scones, recent scientific articles and ethical issues.

Visiting scientists might be more interested in the instruments, such as the 30-centimetre automated telescope with a retractable dome. For this and other expensive equipment, the school is devising a system by which students who have shown themselves to be capable of handling a machine will receive a licence to do so without supervision.

Few schools in the world can match this level of instrumentation. Jim Jarvis, division manager for science and technology at the Thomas Jefferson High School for Science and Technology in Alexandria, Virginia, says that his school, like the YSFH, has a telescope, a scanning electron microscope and PCR machines. But his wish list would include some things that Yokohama has but he does not, such as multiple fume cupboards and gene sequencers.

Judy Scheppler, who directs the Grainger Center for Imagination and Inquiry at the Illinois Mathematics and Science Academy (IMSA) in Aurora, says the nanofabrication and nano-observation facilities at Yokohama are what sets it apart. "Different schools around the country and the world may have some of

YSFH

YSFH

Akiyohshi Wada and Yokahama students (inset). The school (above) is next door to the RIKEN Genomic Sciences Center.

this. But one school having everything is extraordinary," she says.

Sophisticated instruments don't mean much without sophisticated instructors, and the YSFH is providing those too. Before taking up his teaching post, for example, Yutaka Mizogami spent a year training in a Tokyo University laboratory. There he got his name — and his affiliation with the YSFH — in the scientific literature as a co-author[1]. He says he decided to come to the YSFH because he was tired of only being able to give 10% of his time to experiments at his previous school. He is happier now with 30%.

In its first year, the school had more than 5 applicants for every one of its 240 spots for 15–18-year-old students — compared to 3 applicants for a place at the next most popular school in the Kanagawa prefecture. The main reason to come, according to a survey, was the chance to do experiments. In person, some students told *Nature* it was because the teachers are fun. One said he wants to make artificial muscle. Another expressed an interest in methane hydrates.

When talking to students like this, it is hard to believe that there is much to *rika banare*. But politicians have been worrying about science education since 2004, when Japanese 15 year olds dropped from first to sixth in standardized mathematics tests taken the previous year and other science scores started falling. Japan had always prided itself that on such tests it was at or very near the top. By 2006, the last year for which these figures are available, the country ranked tenth in maths and sixth in science.

The counter-attack

Rika banare has inspired the government to designate more than 100 'super science high schools' that receive ¥50 million per year for three years to enhance science education. Wada describes this as a "broad but shallow" first step. Noting the decline in standardized test scores, Wada boasts: "The students at the YSFH will be at the vanguard of a 180° reversal of this trend."

> **"The future of science education in Japan will depend greatly on the success of the school."**
> — Akiyoshi Wada

Wada has long been persuading policy-makers that high-impact science sometimes requires a concentration of resources — although he hasn't always got his way. As a biophysicist in the 1970s, Wada ran into many sceptical biologists when he was one of the first to envision large-scale automated genomic sequencing. But even as these technologies were ramping up elsewhere, Japan's bureaucrats stalled, its genomics fell behind, and when the human genome sequence was finished, Japan accounted for only 6% compared to the 59% and 31% that the United States and United Kingdom produced respectively. The unfolding of Wada's failed efforts are described in a book aptly titled *A Defeat in the Genome Project*[2].

Later, Wada had better luck in turning his vision into reality. He was a key player in creating and maintaining the international Human Frontier Science Program, initiated in Japan in 1989, that has funded 3,000 scientists involved in collaborative projects, including 13 who went on to win Nobel prizes[3]. In 1998, Wada became the founding director of the RIKEN Genomic Sciences Center (GSC) in Yokahama, Japan's first large-scale effort at comprehensive genomics. The generously funded centre led Japan's human- and primate-sequencing efforts and rose to international acclaim with its project to catalogue the active, 'transcribed' parts of the mouse genome[4–5]. "He is a straight-talker, sometimes harsh, but never goes wobbly in his vision and decision," says Yoshihide Hayashizaki, who led the RIKEN mouse project. "I guess that this is why he motivates others to follow him."

Will others follow Wada and his vision for the Yokohama high school? The biggest sticking point in negotiations was the initial cost, which was paid by Yokohama city. Even so, Wada says there have already been half a dozen other regional governments calling to enquire about the school. Scheppler says that select high schools tend to come under pressure because only a few students benefit from their extraordinary facilities. "Shouldn't every student get an excellent science education?" she asks. Wada is sensitive on this point. "Children with outstanding ability, even those from poor households, will be able to take advantage of the low-cost tuition and receive a great education," he says. As it is a public school, students must pay only ¥9,900 a month.

Nobel laureate chemist Ryoji Noyori, president of RIKEN in Wako, says that the school will help to undo the "totally egalitarian public education system in our country". Japan has been chipping away at this way of thinking, and university funding is increasingly focused on competitive funding and centres of excellence. But there is no guarantee that the privileged science track that Yokohama students start on will continue after they graduate. "I fear some students will be disappointed when they later enter the national universities," says Noyori.

Wada is intent on exposing the students to the more luxurious end of Japanese research. The school sits next door to the RIKEN GSC and has collaborative agreements whereby they can use some of the centre's facilities.

Like Wada's other big ventures, this one will eventually be judged on its results: whether the originality being cultured there and at its potential spin-offs translate into improved standardized test scores and a new generation of inspired scientists. Jarvis says that two-thirds of the Jefferson school's 450 annual graduates work in science-related fields. If nanofabrication and PCR can incite a similar passion for science in Yokohama's teenagers, Japan's great experiment in science high schools will have paid off. ■

David Cyranoski is *Nature*'s Asia-Pacific correspondent.

1. Nemoto, T., Watanabe, T., Mizogami, Y., Maruyama, J. & Kitamoto, K. *Appl. Microbiol. Biotechnol.* **82,** 1105–1114. (2009).
2. Ito, Y. *Nature* **433,** 107–108 (2005).
3. Wada, A. *Nature* **357,** 356 (1992).
4. Cyranoski, D. *Nature* **407,** 279 (2000).
5. Kawa, J. *et al. Nature* **409,** 685–690 (2001).

See Editorial, page 151.

CORRESPONDENCE

Interrogation: has abuse been reduced by psychologists?

SIR — Your Editorial 'Responsible interrogation' (*Nature* **459,** 300; 2009), on involvement by psychologists in interrogation at detention centres, was misleading in several respects. You defend psychologists' participation as a protection for detainees, provided that the professional "adheres to, and is held accountable to, the most fundamental medical ethic of all: do no harm".

This ought to be the case. However, there is evidence that psychologists at some US detention centres have been consulting on the use of techniques that amount to torture and cruel, inhumane and degrading treatment, according to the United Nations Convention Against Torture, to which the United States is a signatory.

Psychologists have taught reverse-engineered Survival, Evasion, Resistance, Escape (SERE) techniques to detention-centre personnel. The SERE programme was instigated during the Korean War in order to teach US soldiers to resist torture by the Chinese communists or North Koreans in the event of their capture. To reverse-engineer SERE is to apply the torture that the soldiers were being taught to resist.

Because of these and other abuses, the members of the American Psychological Association approved a referendum on the subject this year prohibiting psychologists' involvement in interrogations at detention centres (see http://tinyurl.com/4xc24r). It is shocking that your Editorial does not mention that this prohibition is the American Psychological Association's policy. By not mentioning it, you leave the impression that psychologists have been helping to reduce abuse, when the data to the contrary are overwhelming.

Frank Summers Psychoanalysts for Social Responsibility, 333 East Ontario, Suite 4509B, Chicago, Illinois 60611, USA
e-mail: franksumphd@hotmail.com

Readers are welcome to comment at http://tinyurl.com/mc3byy

Birds and people both depend on China's wetlands

SIR — Following your News story 'Putting China's wetlands on the map' (*Nature* **458,** 134; 2009) and the related Correspondence 'Time for China to restore its natural wetlands' (*Nature* **459,** 321; 2009), we also wish to stress the need to manage and protect China's existing wetlands, and to remind delegates to the International Congress for Conservation Biology, starting in Beijing on 11 July, of these globally important issues.

Eastern China supports over two million migratory waterbirds outside the breeding season, of which more than one million live in the Yangtze River floodplain. These include concentrations of eight globally threatened species, including the Siberian crane (*Grus leucogeranus*) and the oriental stork (*Ciconia boyciana*), and more than three-quarters of the east Asian populations of seven waterfowl species (L. Cao, M. Barter and G. Lei *Biological Conservation* **141,** 2301–2309; 2008). The wetlands also supply tens of millions of people with food and raw materials, and help to prevent floods and improve water quality.

The influx of water, sediment and nutrients brought by the annual monsoon has maintained these wetlands until now, enabling humans to exploit them without disrupting their spectacular biodiversity. But hydrological changes, especially from hydroelectric and water diversion projects, and the move from low-intensity use by local people to unsustainable exploitation, have severely compromised this long-standing balance.

Safeguarding China's wetland riches requires strategic regional planning, strengthening the nature-reserve network and, most of all, a sound scientific understanding of the processes that support its biological diversity and its productivity.

Lei Cao School of Life Sciences, University of Science and Technology of China, 96 Jinzhai Road, Hefei 230026, Anhui, China
e-mail: caolei@ustc.edu.cn
Anthony D. Fox Department of Wildlife Ecology and Biodiversity, National Environmental Research Institute, University of Aarhus, Grenåvej 14, 8410 Rønde, Denmark

Route for political interests to weaken conservation

SIR — The public often question the validity and economic implications of scientific assessments of biological diversity. Various interest groups, in particular, complain because they weren't consulted. The European Commission therefore decided to open up the assessment process to online commenting, and the outcome will shortly be made public in a report. However, we believe that allowing science and politics to mix is likely to compromise biodiversity conservation.

The European Union's birds and habitats directives form a globally unique legislative framework for protecting and monitoring the status of biological diversity across national borders (see http://tinyurl.com/nk689c). In total, 25 countries have produced 8,820 scientific assessments to evaluate the current status and trends in species populations and habitats of key European interest (see http://tinyurl.com/lz3z8a). One aim of these assessments is to provide relevant scientific information to support the implementation of environmental and sustainable development policies. The information gained will help in setting up new biodiversity goals to replace the targets for 2010, which mostly remain unmet.

The European Commission's decision to include Internet-based comments in the assessment process attracted views from various interest groups and non-specialists with a political agenda. The consultation web tool was used by 170 authors in some 4,000 visits, providing several hundred informal comments on the assessments. The report will reveal the bearing that these have had on the scientific process.

Our concern is that interest groups are now able to manipulate important scientific conclusions. Member states may be persuaded to modify assessments for threatened species such as seals or wolves, for example, because these are perceived to have a harmful impact on populations of economically important species (fish and reindeer, respectively).

It might be argued that such assessments should not be left to scientists, when the political decisions resting on them concern the whole of society. However, policy-makers must base their decisions on accurate scientific information. Allowing inconvenient data to be distorted to guide decision-making could mask important early warning signs.

Science and politics must be kept separate. The task of scientists is to provide transparent information. The task of politicians is to evaluate this information in the light of society's interests before making their decisions. Don't let biodiversity conservation be undermined by political tainting of scientific endeavour.

Petri Ahlroth Finnish Environment Institute, PO Box 140, 00251 Helsinki, Finland
Janne S. Kotiaho Centre of Excellence in Evolutionary Research, Department of Biological and Environmental Science, PO Box 35, University of Jyväskylä, 40014 Jyväskylä, Finland, and Natural History Museum, PO Box 35, University of Jyväskylä
e-mail: janne.kotiaho@jyu.fi

COMMENTARY

Too small to overlook

Voluntary reporting of nanomaterials by industry has failed. Mandatory measures are a step in the right direction, but the field needs more data sharing and oversight, say **Andrew Maynard** and **David Rejeski**.

Global investment in nanotechnology research and development by public and private sectors has risen to more than US$18 billion annually and nanotechnology has passed from a scientific curiosity to a market reality, with hundreds of substances and nanotech-based products now commercially available. Despite the need for evidence-driven oversight of the products of nanotechnology, very little is known about the use of nanotechnology in industry, and relevant oversight remains in its infancy. Even basic information on what nanomaterials are being produced, in what quantities and how they are being handled and used, remains largely undocumented.

Voluntary efforts to gather information on the production of nanomaterials have been largely unsuccessful. A UK reporting scheme initiated in 2006 received just 13 submissions in two years. A similar US Environmental Protection Agency (EPA) programme did not fare much better. It had received only 29 submissions by the end of 2008, covering 123 different nanoscale materials. The agency's own assessment concluded "it appears that approximately 90% of the different nanoscale materials that are likely to be commercially available were not reported"[1]. Low reporting levels are probably accounted for by the requirements for businesses to expend considerable resources in gathering information, with little direct benefit to themselves.

Against this backdrop, in January, Canada announced that it would initiate a one-off mandatory data collection from industry on the production and importation of engineered nanomaterials in quantities of greater than 1 kilogram, to help inform future regulatory decision-making. France is also in the process of introducing mandatory reporting — a move that could put it at odds with its European neighbours by creating a different regulatory environment.

Instituting mandatory reporting is a welcome move towards effective oversight, but it is only one of many steps that are needed if the promise of this technology is to be realized.

In principle, such data calls and any resulting evidence-based regulations will benefit industry by reducing uncertainty and by clarifying operational guidelines. Solid data on what nanomaterials are being produced and used (and in what quantities), what the potential risks are and how adverse effects can be minimized or avoided, will enable effective risk management. The dearth of information on human health and environmental effects is being noticed by insurance companies — for example, Lloyd's of London has placed nanotechnology in its top tier of emerging risks — raising concerns that a continued lack of data could increase the cost of company insurance.

U. BELLHAEUSER/GETTY IMAGES

The risks posed by products made with nanotechnology need to be quantified.

Effective oversight will depend on effective data sharing and on targeted research. Five research challenges that would help support evidence-based decisions on using nanotechnology safely were suggested in *Nature* in 2006 and they are still valid today[2]. These include exposure monitoring, toxicity testing, predicting and avoiding harmful behaviour and evaluating risks for the lifespan of the material. Several countries are addressing these challenges, but research is about 2–5 years behind the proposed timetable, and information gaps still exist in how to generate, use and dispose of engineered nanomaterials safely[3].

Increase the flow

For research and regulation to work, more needs to be done to facilitate the flow of information between organizations. New government and industry partnerships are needed to ensure access to relevant and trusted data on nanomaterial risks. Small firms and start-up companies need help from government and industry associations to address potential risks and meet regulatory requirements. Mechanisms also need to be established to enable industry and researchers to share data with regulators, consumers and other businesses.

Already, non-governmental initiatives such as the GoodNanoGuide[4] and the Responsible NanoCode[5] are helping businesses work together to address nanomaterial safety. National and international standards organizations are developing guidance on how to work safely with nanomaterials. Additionally, the Organisation for Economic Co-operation and Development is coordinating efforts to generate toxicity data on 14 commercially relevant nanomaterials that will be publicly available.

Regulators are also acting, even in the absence of data. On 29 June, the EPA tightened its regulatory oversight on specific types of carbon nanotubes. Later this year, European rules governing the use of nanomaterials in cosmetics will come into effect, and the application to nanomaterials of European regulations on chemicals continues to be explored.

Yet progress remains patchy, and developers and regulators are a long way from understanding how to predict and manage the potential risks associated with existing nanomaterials, let alone those in the pipeline. Moves towards mandatory data collection are welcome, and we hope other nations will follow suit. But it must be accompanied by more effective and timely data sharing and oversight at the local, national and international level. Get this right, and consumers and industry stand to gain from the development of safe, sustainable new technologies. But get it wrong, and everyone is likely to lose out. ■

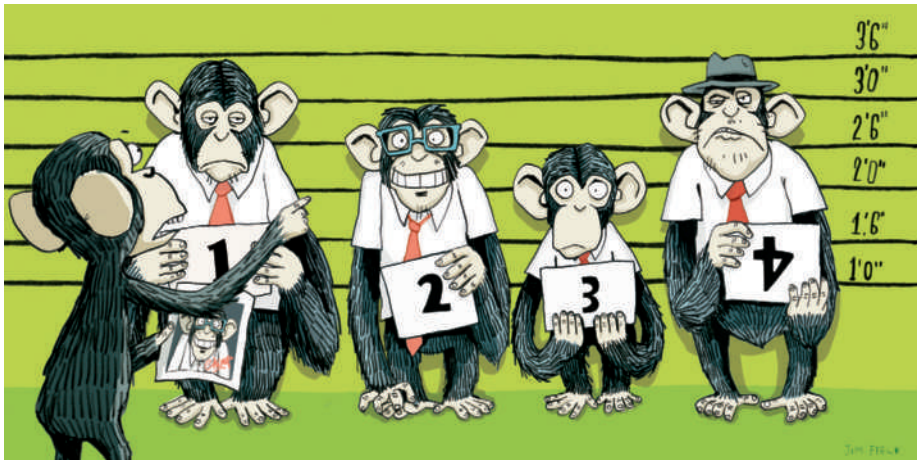
Andrew Maynard and **David Rejeski** are at the Project on Emerging Nanotechnologies at the Woodrow Wilson International Center for Scholars, 1300 Pennsylvania Avenue, Washington DC 20004-3027, USA.
e-mail: andrew.maynard@wilsoncenter.org

1. *Nanoscale Materials Stewardship Program. Interim Report* (US Environmental Protection Agency, 2009).
2. Maynard, A. D. *et al. Nature* **444,** 267–269 (2006).
3. *Risk Assessment of Products of Nanotechnologies* (Scientific Committee on Emerging and Newly Identified Health Risks, 2009).
4. www.goodnanoguide.org
5. www.responsiblenanocode.org

ESSAY

Darwin's last laugh

We must look for mental commonalities between humans and other animals to understand the minds of either, says **Frans B. M. de Waal**, rebutting a recent claim to the contrary.



JIM FIELD

In 1739, the Scottish philosopher David Hume wrote: "When any hypothesis … is advanc'd to explain a mental operation, which is common to men and beasts, we must apply the same hypothesis to both." A century later, Darwin showed that all forms of life have a common origin. Yet, to this day, the idea that humans and animals share characteristics and abilities, including mental ones, as a result of shared evolutionary history, still seems hard to swallow for some.

For example, in a recent critique of evolutionary approaches to cognition[1], Johan Bolhuis and Clive Wynne label the anthropomorphism of Charles Darwin "farfetched". They question those who, like Darwin, believe that there is "no fundamental difference between man and the higher mammals in their mental faculties". Attempts to identify human-like cognition in other animals has invariably led to over-interpretation, they argue.

I disagree. The opposite approach of anthropodenial — the *a priori* rejection of continuity between humans and other animals — has led people to systematically underestimate animals[2]. Well into the last century, comparative psychologists had animals perform arbitrary laboratory tasks unrelated to the problems they face in their natural environments. This theory-free 'behaviourism' never advanced our understanding of cognition to the degree that Darwinism has.

> **"Efforts to single out behaviours as distinctly human have rarely held up to scrutiny."**

Evolutionary theory predicts cognitive similarities based on the relations between species and their habitats. It also tells us that if closely related species, be they octopus and squid or human and chimpanzee, show similar responses under similar circumstances, the most parsimonious interpretation is that the cognition involved is similar too. Humans and their closest relatives diverged so recently, in evolutionary terms, that it is hardly anthropomorphic to assume that shared ancestry suggests shared cognition.

A wealth of recent evidence supports this assertion, most of it discovered precisely because investigators have taken human capacities as their starting point. Only humans were thought capable of recognizing faces from the arrangement of the nose, eyes, mouth and so on. But other primates have this ability, and the same neural substrate seems to be involved[3]. Similarly, bonobos, golden monkeys and a variety of social mammals kiss, embrace, groom or mount their opponents after a fight. Calling this 'reconciliation', a term derived from human interaction, has proven appropriate given that these reunions alleviate stress and repair social ties[4]. In contrast, efforts to single out distinctly human capacities have rarely held up to scientific scrutiny for more than a decade, such as claims about culture, imitation, planning and the ability to adopt another's point of view.

Other behaviours may have an even longer evolutionary history. For example, a nasal spray of oxytocin, a hormone and neurotransmitter common to all mammals, can enhance a person's tendency to share money with another. A related hormone, vasopressin, is known to strengthen pair bonds in rodents, and the effect of oxytocin on non-human primates is being tested.

Even distantly related species, such as elephants, dolphins, primates and birds, share an evolutionary history that may explain cognitive similarities, much as deep homologies in genetic instruction underlie the eyes and limbs of both flies and rodents. For example, neuroscientists first discovered mirror neurons in macaques, but have since found them in swamp sparrows, suggesting that they occurred in the common ancestor of birds and mammals. These neurons fire both when an animal performs an action and when it sees or hears another perform that action, and are thought to facilitate human imitation and empathy.

Some behavioural similarities will be the result of convergent evolution, in which species evolve similar cognitive capacities independently, because they have been exposed to similar selection pressures. For example, birds that store their food, such as scrub jays, need to know when competitors can see them. They use deceptive tactics akin to those of chimpanzees and other primates that live in large groups[5]. Likewise, capuchin monkeys and Caledonian crows, with similar foraging needs, have both ended up using tools. Yet even here, we cannot rule out the potential role of shared ancestry, given that bird and mammal brains are not nearly as different as once thought.

In sum, there is no good scientific reason to give evolutionary approaches short shrift, or to deride Darwin's speculations about continuity between humans and other animals, including a "sense of humour" — even the playful panting sounds of apes have recently been shown to be homologous to human laughter. Anyone who has watched primates, elephants or ravens at play realizes that here, too, Darwin may well have the last laugh. ■

Frans B. M. de Waal is director of the Living Links Center, part of the Yerkes National Primate Research Center, Emory University, 954 N. Gatewood Road, Atlanta, Georgia 30322, USA, and is author of the forthcoming *The Age of Empathy* (Harmony, 2009).
e-mail: dewaal@emory.edu

1. Bolhuis, J. J. & Wynne, C. D. L. *Nature* **458,** 832–833 (2009).
2. de Waal, F. B. M. *Phil. Top.* **27,** 255–280 (1999).
3. Parr, L. A. *et al. Curr. Biol.* **19,** 50–53 (2008).
4. de Waal, F. B. M. *Science* **289,** 586–590 (2000).
5. Emery, N. J. & Clayton, N. S. *Science* **306,** 1903–1907 (2004).

BOOKS & ARTS

Orderly anarchists

The profit motive has led pirates to come up with surprisingly democratic and egalitarian social structures. It is a lesson in bottom-up economics, explains **Michael Shermer**.

The Invisible Hook: The Hidden Economics of Pirates
by Peter T. Leeson
Princeton University Press: 2009.
271 pp. $24.95

Do modern-day pirates, such as these suspected Somali ones, need order and honesty to turn a profit?

HO/REUTERS/CORBIS

In recent years, economists have joined the ranks of high-profile scientist writers in publishing thoughtful books intended for both the general public and their colleagues. In works such as *Freakonomics* by Steven Levitt and Stephen Dubner and *Nudge* by Richard Thaler and Cass Sunstein, seemingly every aspect of human life is examined from an economic perspective. Peter Leeson's book is a good addition to the genre.

A major theme of such books is that behaviour that seems irrational is in fact rational when economic incentives are considered. Take piracy. In *The Invisible Hook*, Leeson, an economist at George Mason University in Fairfax, Virginia, argues that acts such as flying the Jolly Roger flag, concocting code rules, and employing forms of punishment are rational responses to the pursuit of profits. Invoking Adam Smith's powerful economic metaphor of the "invisible hand", Leeson lays bare the structure of pirate societies. Along the way he offers one of the finest introductory courses in economics since Henry Hazlitt's *Economics in One Lesson*.

The public, Leeson explains, view pirates as "liars, cheaters, and traitors" and pirate society "as orderly and honest as an asylum for the criminally insane … without a warden". This attitude is older than the *Pirates of the Caribbean* films: in 1726, for example, King George I received a petition from "the General Officers of the Army" that pirates were "profess'd enemys to all Order and Government".

This perception is wrong, according to Leeson. No community can succeed if it is utterly anarchistic. Adam Smith recognized this point a half-century after George I: "If there is any society among robbers and murderers, they must at least … abstain from robbing and murdering one another." Leeson argues that pirate life had to be "orderly and honest" to meet its economic goals. These were those of any corporate enterprise: to turn a profit.

Lacking the social structures and political tools of civil society, pirates invented their own. *The Invisible Hook* shows what pirate order looked like, how it worked, and the incentives needed to maintain it even as they disrupted order on the sea. It is also a lesson on how social structure forms naturally from the bottom up out of economic necessity, instead of from the top down by political fiat. Just as it has been shown that the Wild West of nineteenth-century America was a relatively ordered society, in which ranchers, farmers and miners came up with their own rules and institutions for conflict resolution long before the federal law could reach them, Leeson reconstructs from historical documents how pirate communities did the same thing. Benjamin Franklin allegedly said at the signing of the Declaration of Independence: "We must, indeed, all hang together, or assuredly we shall all hang separately." Surely some pirate captain conveyed the same sentiments to his crew, perhaps as a warship was bearing down upon them with guns blazing.

Pirates employed forward-looking economic practices. Many pirate ships elected their captains and had a strict set of rules for everyone to follow, placing restrictions on problem activities such as drinking, gambling, sex, desertion and fighting. Pirate sailors were better paid than those in many marine navies, they were more tolerant of racial diversity among crews, employed clearer systems of corporal punishment, and divided the spoils with greater equanimity than their naval counterparts. Shirking one's duties during battle was particularly worthy of punishment because it could lead to the "free-rider" problem where loot was divided evenly between uneven efforts, breeding resentment, retaliation and "an-arrgh-chy". Pirate codes were specific in their laws and punishments for breaking those laws, to which pirate crews had to consent before sailing.

Leeson tracked down the sharing of contractual arrangements between captains, made possible by the fact that "more than 70 percent of Anglo-American pirates active between 1716 and 1726, for example, can be connected back to one of three pirate captains", and thus the pirate code emerged from "piratical interactions and information sharing" not from one central pirate king.

Across history, merchant and military navies have themselves hardly been beacons of enlightened liberalism, engaging as they did in questionable practices such as the British,

French, and Spanish pillaging of native American resources in the sixteenth and seventeenth centuries, during the heyday of piracy.

Sovereign governments may have legalized such plundering, but they were not necessarily more moral than the pirates who re-plundered that same wealth. Both used the threat of force, as Leeson reminds us. He does not argue for moral equivalency, rather he explains that pirates form their own versions of civil societies for the same reason everyone else does: economic success.

The Invisible Hook is a good addition to the genre of popular economics: a fun and enlightening read, and rock solid in its scholarly *bona fides*. ■

Michael Shermer is publisher of *Skeptic* magazine, a columnist for *Scientific American*, and author of *The Mind of the Market*.
e-mail: shermer@skeptic.com

Solo journey to a fifth dimension

Hypermusic Prologue: A Projective Opera in Seven Planes
Hèctor Parra and Lisa Randall
Pompidou Centre, Paris. 14–15 June 2009

An opera about string theory and five-dimensional space is hard to imagine. But one premiered recently in Paris.

Hypermusic Prologue is a collaboration between composer Hèctor Parra and Lisa Randall, professor of theoretical physics at Harvard University and author of *Warped Passages: Unravelling the Universe's Hidden Dimensions* (Allen Lane, 2005), an account of cutting-edge physics, including string theory, and the possibility of additional spatial dimensions beyond those we sense. Singers James Bobby and Charlotte Ellett sang Randall's libretto, accompanied by musicians and technicians of the Paris-based Ensemble Intercontemporain. All gave admirable performances, with flashes of startling beauty.

Parra says that when he first read *Warped Passages*, something clicked. "For me it was a real discovery," he said. "For a few months I imagined reality vibrating like strings." The composer, who studied engineering before choosing music and whose father is a physicist, imagined a chamber opera in which music represented the phenomena of high-energy physics. He contacted Randall, who was keen to try a new genre. "In a book, if someone gets to some idea that they don't understand, they will stop", she explains. "You have more licence in a dramatic performance."

Parra's music is experimental, filled with glittering, jarring, liquid, fractured-icicle sounds, often electronically generated. It is as far from classical harmony as string theory is from classical mechanics. "It was the right kind of music for this kind of physics precisely because it is esoteric," says Randall.

The plot of *Hypermusic* is simple: a pair of lovers, played by Ellett and Bobby, separate because the woman feels something is missing, both in her life and in physics theory. Both characters are physicists; the female protagonist is also a composer. In the middle of the hour-long show, she takes a trip into the fifth dimension, which she experiences with vivid delight. "It is a nice metaphor for exploring a new world," says Randall. "Of course, it's unrealistic." Also unrealistic is that the male character seems to be intellectually stuck in pre-Einstein physics. Randall added that "just for humour value".

DR; L. HOSSEPIED

Lisa Randall, physicist and librettist of *Hypermusic Prologue*; opera set above.

Psychological depth gets intentionally short shrift. "The two characters are schemes, sketches," Parra explains. And the quantity of technical language in the libretto is breathtaking. "There's more physics in it than I had intended," says Randall. The composer and the production designer convinced her to include more, for atmosphere. "When you watch a movie about a painter you don't learn how to paint," she explains. "But you learn what it was for the painter to get immersed into that world."

Randall describes in *Warped Passages* how our familiar world is confined to a four-dimensional space-time 'brane' that lies, in her theory, within a larger five-dimensional 'hyperspace'. Moving into the fifth dimension takes the fictional traveller into regions of vastly magnified gravity that distorts other attributes of reality and experience: time, distance, energy and mass.

The challenge was to depict this exotic journey as a beautiful experience for the audience. Parra samples the sounds produced by the singers and instruments and passes them through an elaborate digital system of real-time signal processing and synthesis. The instrumental and vocal scores are of stunning complexity, with more than 100 parameters of digital transformation, which evolve as the plot progresses. The warping of physical time, as experienced by Ellett's character, for example, is expressed through modulations of rhythm and the 'granularity' of the synthesized music.

The processed sounds are projected into the auditorium from an array of speakers, using signal delays that constantly shift the apparent locations of their sources. The audience hears this complex aural texture blended with the sounds produced directly by singers and instrumentalists. "You hear 70% the real musicians and 30% electronics, on average," Parra says.

The set is an articulated screen on which artist Matthew Ritchie projects video images evoking the two characters' conflicting perspectives. The kaleidoscopic, lava-lamp style of the images evokes a psychedelic atmosphere that seems slightly out of register with the idea that the fifth dimension is physically real.

By contrast, Lisa Randall is a scientist who stays close to experimental data. "I'm not a string theorist, I'm a model-builder," she says. "One of my fears was that by the time this happened we'd find out it wasn't even right." Experiments testing her ideas are expected to run in the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, after it starts up again later this year. With luck, they might coincide with upcoming performances of *Hypermusic*. "It's a good time for science to inform art," says Randall. "People are interested. If there were extra dimensions, people would want to know." ■

Stefan Michalowski is a former particle physicist, currently working on intergovernmental science policy; **Georgia Smith** writes for the *International Herald Tribune*'s 'Ear for Opera' series.
e-mails: stefanm@noos.fr; georgias@noos.fr

***Hypermusic Prologue* will be performed on 27–28 November at Gran Teatre del Liceu, Barcelona, Spain, and on 6 December in the Grand Auditorium of the Philharmonie, Luxembourg.**

Where they lived

Lived in London: Blue Plaques and the Stories Behind Them
Edited by Emily Cole
Yale University Press: 2009. 637 pp.
$85, £40

When Einstein died in 1955, *The Washington Post* published a cartoon that has since become famous. It shows Earth floating among the other planets and heavenly bodies, with a sign tacked on to it bearing the words "ALBERT EINSTEIN LIVED HERE".

Famous past residents of London are similarly recalled by a series of circular blue plaques that adorn their one-time mansions, houses and flats. The scheme has been imitated by major cities around the world. Their stories are told in a new complete illustrated guide, *Lived in London,* which documents recipients street by street, square by square. Yet, less than a tenth are scientists. Of around 800 people who are commemorated with plaques, 58 are scientists, along with 44 medics, 17 engineers and 15 others associated with industry and invention. This suggests some bias against recognizing scientists beyond a charmed circle of celebrities — there are 194 plaques for literary figures.

Significantly, the first scientist to be honoured was Benjamin Franklin, who was a prominent statesman as well as a prominent scientist. During Franklin's sojourns in London, in 1757–62 and 1764–75, accompanied by his son and their two slaves, he was the agent of the Pennsylvania Assembly, charged with encouraging pro-American sympathies in the run-up to the Declaration of Independence.

Isaac Newton, like Franklin, lived in the city for a long period but did his best scientific work elsewhere. Michael Faraday, Christopher Wren and Norman Lockyer (the founding editor of *Nature*) spent almost all of their working lives in London. Others, such as Charles Darwin and the Italian Guglielmo Marconi, made breakthroughs there, but stayed only a short time. Darwin discovered the principle of natural selection while living in Bloomsbury in 1838–42.

The criteria for choosing an individual have inevitably varied since the official system of commemorative plaques, now run by English Heritage, was started by the then Society of Arts (now the Royal Society of Arts) in 1866. Certain figures, such as Thomas Henry Huxley, were commemorated soon after they died. Nowadays a blue plaque can only be awarded at least 20 years after a person's death. But the main criteria have remained the same. Individuals must be regarded as eminent by fellow professionals and to have made "an important positive contribution to human welfare or happiness". Place of birth or nationality is irrelevant, but they must have resided in London for a significant period — in time or in importance. Finally, either the "well-informed passer-by" must recognize their name, or they must deserve national recognition. A varied list results.

D. KENDALL/ENGLISH HERITAGE

Isaac Newton's former Jermyn Street, London address.

Despite Henry Cavendish being little known to the public, the house of this great eighteenth-century experimental physicist is marked in Bloomsbury's Bedford Square. Here he established his collection of minerals and a semi-public library of some 12,000 volumes. His private laboratory was at Clapham Common in south London. Cavendish, the book reports, was "taciturn to the point of Trappism", and communicated with his three servants by note.

H. G. Wells, by contrast, had massive name recognition but lack of approbation from scientists. By the time he moved into his house near Regent's Park in 1936, Wells was one of the most famous writers in the world, partly through his science fiction, such as *The Time Machine*, partly through best-selling non-fiction such as *The Outline of History*. Despite his claim that his house was "tumble-down" in a self-composed obituary, its luxuries included a four-poster bed and a private telephone exchange. Wells never received the Fellowship of the Royal Society that he felt was his due, and publicly referred to the fellows as "a lot of bastards".

Between these two poles, there is much scope for differences of opinion. Physicist and inventor Charles Wheatstone, who died in 1875, was initially rejected for a blue plaque in 1906 despite his contributions to telegraphy and the measurement of electrical resistance — using the 'Wheatstone bridge'. In the end, the proposal was successful, revived not by a scientist but by a member of the International Concertina Association, who drew attention to Wheatstone's invention of the instrument, which he patented in 1844.

Lived in London groups the plaques by borough or district, rather than by alphabetical order of recipient, and includes short biographies. These are a well-judged mixture of facts, achievements, anecdotes and quotations, covering the person's entire life, not just their time in London. Of the 'names' with whom I am most familiar, the nineteenth-century physicist, physician and polymath Thomas Young has an exemplary entry, but the book misses the fact that architect and archaeological decipherer Michael Ventris deciphered Minoan Linear B in a flat in Highgate before he moved to the house in neighbouring Hampstead to which his plaque is attached.

Einstein, alas, never lived in London, despite making influential visits there. So there can be no plaque for him. But for others who stayed longer in the city, the blue plaques offer a simple but effective way of bringing past scientific achievements alive to inhabitants and visitors alike. ■

Andrew Robinson is the author of *The Last Man Who Knew Everything*, a biography of Thomas Young, and *The Man Who Deciphered Linear B*, a biography of Michael Ventris.
e-mail: ar471@cam.ac.uk

Fresh formulae for portraiture

New portraits of physicists David Brewster and Peter Higgs show that naturalistic images can find distinct ways to reflect scientists and their work, **Martin Kemp** explains.

Many portraits depict scientists just as they are, with nothing to tell us what the sitter actually does. This was the dominant mode for centuries, when the key ambition was to be shown as a 'gentleman'. Props could also be imported — a flask or telescope, say — to give a clue about the individual's field of excellence. Then, from the nineteenth century, particularly in photography, scientists began to be shown in their working habitats — sometimes even in action.

But these formulae have become tired. Can an artist find a way of making the format or nature of a portrait stand in a more imaginative relationship to the achievements and character of the individual scientist?

An extreme solution was presented by Marc Quinn's 'genomic portrait' of John Sulston, composed from the cultivation of a fraction of Sulston's own DNA, which involves no traditional 'likeness' at all (M. Kemp *Nature* **421,** 416–420; 2003). Alternative solutions to the dilemma are posed by new portraits unveiled last year of two Scottish university scientists, one past and one present.

The historic image portrays David Brewster, a combative nineteenth-century authority on optics, biographer of Newton, pioneer of photography, and inventor of the kaleidoscope and lenticular stereoscope. It is by Calum Colvin, who specializes in his own kind of optical magic, and was displayed earlier this year at exhibition *Natural Magic* at the Royal Scottish Academy in Edinburgh.

Colvin characteristically creates large-scale colour photographs of elaborately staged set-ups. He assembles tableaux of objects, such as tables, chairs and smaller domestic items, much like a seventeenth-century Dutch painter of interiors. He then selectively covers their surfaces with painted marks in such a way that they assemble themselves into a quite different image when photographed from a specific viewpoint. On angled and curved surfaces, the technique required is that of anamorphosis — the trick of oblique perspective that was perfected during the sixteenth and seventeenth centuries.

The Brewster portrait is one of a series that extends Colvin's illusions into the arena of stereoscopy, which exploits our binocular vision to create a three-dimensional effect from two-dimensional images. The series includes a portrait of Charles Wheatstone, Brewster's great rival and adversary, the inventor of the stereoscopic viewer.

Portraits of theoretical physicist Peter Higgs (above) and kaleidoscope inventor David Brewster.

Resorting to the stereo technique simultaneously clarifies the three-dimensional reality of Colvin's arrays and masks the physical reality of the photograph itself as a surface. As he says, "I wanted to remove this physicality, making the images a form of 'mental construct'."

Looking through a lenticular viewer at two images of the Brewster portrait tableau photographed from separated viewpoints, we make out a ladder — for Colvin a symbol of intellectual ambition — and a clutter of objects specific to Brewster. The latter includes a stereo photograph of John Adamson, chemist and practitioner of photography in the Scottish town of St Andrews when Brewster lived there. In 1832, Brewster published his *Letters on Natural Magic*, addressed to the novelist and poet Walter Scott: Colvin has used Brewsterian magic to blend subject and scientific activity in a remarkable way.

The living scientist is Peter Higgs, the University of Edinburgh theoretical physicist famed for the boson that bears his name, devised to explain the mass of elementary particles. The search for the Higgs boson is one of the focuses of the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, which is scheduled for relaunch in autumn, and is now looking over its shoulder at Fermilab in Batavia, Illinois. The painter of Higgs's portrait is Ken Currie.

The painting is set in Higgs's flat, one typical of Edinburgh's New Town district, which Currie describes as having "a very particular ambience — a greyish light illuminating the paraphernalia of his life". Among the paraphernalia are portraits of grandchildren below a massive mirror that is an original feature of the flat. The mirror itself is viewed from a standpoint equivalent to some 3 feet from the right edge of the canvas. In it we witness radiant traces of "the often very elegant movements of colliding particles".

Higgs, with his spectacles in hand behind his back and his eyes located precisely along the horizontal of the perspective, seems to be staring simultaneously at us and, in reflection, at the 'final proof', which exists in an undefined space on our side of the canvas. Currie plays on the contrast between the seeping natural light and the febrile brilliance contrived in the recording instruments of the collider. The one light is earthly; the other emanates from a "world that is sublime, awesome, strangely terrifying", in the painter's own words. A quotation from the seventeenth-century thinker Blaise Pascal haunted Currie's imagination while he was painting the portrait: "The eternal silence of these infinite spaces fills me with dread."

A sheet of paper and pen are tucked into the standing physicist's pocket, in contrast to the awesome technology deployed in the collider's search for elusive boson, buried in realms of the scientific sublime. This primitive equipment is, as Higgs told Currie, what he needs for his work as a researcher of pre-experimental concepts. In a sense the portrait presents us with this quest within Higgs's mind. It is suffused by the longing for discovery and proof that drives all scientific thought, whether theoretical or experimental.

During the twentieth century, portraiture of all kinds became deeply problematic, as the great majority of experimental artists moved away from naturalistic representation. There was reason to think that the painted portrait was dying as a major genre. What both these artists show, faced with the challenge of portraying scientists who are involved with different modes of seeing and visualization, is that the traditional portrait, like the novel, can be endlessly revitalized when form and content are reforged into new kinds of unity. ■

Martin Kemp is emeritus professor in history of art at the University of Oxford, Oxford, UK.

'PORTRAIT OF SIR DAVID BREWSTER' 2008 CALUM COLVIN; K. CURRIE, FLOWERS GALLERY LONDON

NEWS & VIEWS

EVOLUTIONARY BIOLOGY

Microbes exploit groundhog day

Tim F. Cooper

Can microorganisms learn from history? When a sequence of environmental changes is repeated, natural selection might select for responses that enable the microbes to prepare for later challenges in the sequence.

Ivan Pavlov is famous for demonstrating that dogs can learn to associate events that occur close together in time[1]. By repeatedly preceding food delivery with an unrelated stimulus (sometimes bells, sometimes electric shocks), Pavlov was able to condition dogs to link these events. After a while, the stimulus was sufficient for the dogs to anticipate food. On page 220 of this issue, Mitchell *et al.*[2] present evidence that microorganisms can benefit from conditioned responses that allow them to prepare for environmental changes.

The proposal that microorganisms can associate a stimulus with an appropriate response to a future environment might seem far-fetched. After all, without cognition, microorganisms rely on simple regulatory networks to sense and respond to their environment. A canonical example of gene regulation, the response of *Escherichia coli* to the sugar lactose, illustrates why it seems surprising that such networks can be used to anticipate environmental changes. This bacterium activates genes needed for lactose use (the response) only in the presence of lactose or similar chemicals (the stimulus). The response is mediated by a specific interaction between lactose and a regulatory protein that lessens the activity of the protein, reducing its ability to repress the lactose genes. This direct link between the stimulus and the response makes sense: it reduces costly gene activation in the absence of lactose, while ensuring that activation occurs when it is useful.

Insight

If regulatory networks function to sense the current environment, how is it possible for them to induce a response that will be beneficial only in a future environment? The insight of Mitchell *et al.*[2], building on previous work[3], was to realize that the connection between stimulus and response can be offset in time. For example, if a non-lactose sugar consistently follows the availability of lactose, selection might favour the evolution of a regulatory network that directly links the presence of lactose to the expression of the non-lactose-utilization genes. This network would serve to 'prime' cells, conferring an advantage by preparing them to use the non-lactose sugar in anticipation of its imminent availability, and thereby reducing the lag time characteristic of *de novo* activation of response genes. Mitchell *et al.* call this mechanism adaptive anticipatory conditioning.

But how can this idea be addressed experimentally? Mitchell *et al.* examine the responses of *E. coli* and baker's yeast, *Saccharomyces cerevisiae*, to sequences of environmental changes that it is reasonable to think they may have been repeatedly exposed to throughout their evolutionary histories. For *E. coli*, the authors consider stimuli and appropriate responses associated with passage through the initial region of the gastrointestinal tract (an environment relatively high in lactose but low in the sugar maltose), and potentially adaptive responses in the lower part of the tract (an environment relatively low in lactose and high in maltose). For yeast, they consider the temporal sequence of stresses associated with fermentation: high temperature, increasing levels of ethanol and exposure to oxidative stress.

For both *E. coli* and yeast, Mitchell *et al.* find evidence that early environmental stimuli prime cells to better deal with later environments. In *E. coli*, lactose (the early stimulus) caused intermediate activation of maltose genes (the late response). This priming provided a growth benefit, relative to unprimed cells, when cells were subsequently transferred to an environment containing maltose. Similarly, in yeast, early stress stimuli induced responses that conferred protection to later stresses. For example, pre-exposure to ethanol conferred a 29-fold increase in survival to subsequent exposure to an oxidative stress that simulated late-fermentation conditions.

Priming alone, however, is not sufficient to infer that an anticipatory response has taken place. For example, it is possible that activation of maltose genes may provide *E. coli* with a benefit during growth in lactose. If so, benefits of priming following transfer to a subsequent environment could be a side effect of this direct advantage. To address this possibility, Mitchell *et al.* took advantage of an experiment in which *E. coli* were evolved in an environment containing only lactose[4]. In this environment, the temporal link between the presence of lactose and future maltose availability is removed. Consistent with the priming response being selected to prepare cells for future availability of maltose, the presence of lactose no longer induced maltose genes in the evolved strains. Moreover, this change was associated with the loss of the lactose pre-exposure advantage before transfer to maltose.

Responses

Mitchell and colleagues' experiments[2] show that microorganisms can interpret their environment and respond in a way that provides a benefit only following a future environmental change. The ecological forces, if any, selecting for this ability are less clear. Responses that seem anticipatory in carefully defined lab environments may simply be side effects of imperfect regulation, providing no benefit in complex natural environments. Alternatively, the responses could provide a direct benefit in an unknown but important environment relevant to selection in nature. Collection of detailed fitness and ecological data will be required to address these points. What are the relative costs and benefits of being primed in natural environments? How reliably does a potential stimulus precede a particular environmental change? Depending on the answers to these questions, it may turn out that other response strategies are more successful.

One message is clear. The regulatory networks that link environmental stimuli to microbial responses are complex and can evolve rapidly. The potential for microorganisms to offset responses from environments in which those responses are useful provides both a warning and an opportunity for researchers involved in testing the functional significance of links between stimuli and responses. ■

Tim F. Cooper is in the Evolution and Ecology Group, Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204, USA.
e-mail: tfcooper@uh.edu

1. Pavlov, I. P. *Conditioned Reflexes* (Dover, 1960).
2. Mitchell, A. *et al. Nature* **460,** 220–224 (2009).
3. Tagkopoulos, I., Liu, Y.-C. & Tavazoie, S. *Science* **320,** 1313–1317 (2008).
4. Oxman, E., Alon, U. & Dekel, E. *Evolution* **62,** 1547–1554 (2008).

VISION

New light on allergy receptor

Maria Grant

A receptor usually found on immune cells implicated in allergy turns out to be a diagnostic marker and promising treatment target for a degenerative eye disease. Curiously, its role in the eye seems to be unrelated to inflammation.

Age-related macular degeneration (AMD) affects millions of people worldwide[1] and is the leading cause of blindness in elderly people in the United States[2]. Treatment options for AMD are limited, and people are usually diagnosed late in the course of the disease, by which time they often have permanent vision loss. However, on page 225 of this issue, a study by Takeda and colleagues[3] may usher in a new era in the management of this condition. They report the identification of a biological marker for early AMD that could be exploited for timely diagnosis and treatment.

There are two forms of AMD — 'wet' and 'dry'. Although the dry form is more common, it is less likely to cause severe bilateral visual loss. In the wet form, abnormal blood vessels develop from the choroid — the tissue that lies external to the sensory retina — in a process called choroidal neovascularization (CNV). The new choroidal vessels bleed or leak plasma proteins into the retina, leading to retinal damage. This particularly affects the macula region that resolves fine visual detail, with loss of central vision that results in blindness.

In the past, CNV has been treated with photodynamic therapy (a form of treatment in which laser activation of a drug results in occlusion of blood vessels) or surgery. Recently, biological therapy has become available for clinical use. Treatment consists of agents that block the activity of a protein that induces the formation of new blood vessels, known as vascular endothelial growth factor-A (VEGF-A)[4]. Although such treatment has not been found to harm the retina substantially in humans, concerns have been raised about its long-term safety because VEGF-A is also involved in the growth of normal blood vessels, and in the maintenance and function of neuronal cells in the adult retina[5]. In addition, not all patients respond adequately to blocking of VEGF-A.

Takeda and colleagues[3] now identify a new molecular player in AMD pathogenesis. They studied mouse models of AMD and found that intraocular expression of CCR3 — a receptor found on eosinophils and mast cells, which are immune mediators of allergic inflammation — is essential for the growth of choroidal vessels in CNV. CCR3 is a G-protein-coupled receptor that binds to several small chemoattractant proteins, known as CC-type chemokines. These include members of the eotaxin family — eotaxin-1 (also known as CCL11), eotaxin-2 (CCL24) and eotaxin-3 (CCL26) — which mainly recruit eosinophils to sites of inflammation, and are known to have a role in the pathogenesis of asthma and allergy[6–8].

When Takeda *et al.*[3] blocked intraocular CCR3 with neutralizing antibodies or receptor antagonists in a mouse model of CNV, they found a significant decrease in the generation of abnormal blood vessels in the choroid after laser injury. CCR3-neutralizing antibodies inhibited laser-induced CNV more effectively than VEGF-A-neutralizing antibodies. The same was true for knockout mice that lack CCR3. In these mice, the amounts of eotaxin-1 and eotaxin-2 in the retinal pigmented epithelium (the layer of cells that lies between the choroid and the retina) were markedly increased after laser injury, and the authors also showed these chemokines to be crucial for CNV development.

Importantly, the authors extend their finding to humans. They show that CCR3 is expressed only in the endothelial cells lining the abnormal blood vessels of CNV from people with wet AMD and not in choroidal endothelium from people without AMD, or from those with dry AMD or other types of choroidal or retinal disorder. These pivotal observations demonstrate that CCR3 is a biomarker that is specific for both proliferating choroidal endothelial cells and wet AMD. Takeda *et al.* verify that eotaxin-1, -2 and -3 are expressed in human choroidal neovascular tissue from people with AMD, suggesting that the eotaxin–CCR3 axis is functional in human disease (Fig. 1).

Intriguingly, CCR3 and the eotaxins do not seem to contribute to the pathogenesis of AMD through the recruitment of eosinophils or mast cells, as neither type of cell is found in human CNV. Also, blocking of CCR3 reduced CNV in mice that were rendered deficient for either eosinophils or mast cells. The recruitment of neutrophils and macrophages, cells that have been shown to be crucial for the development of experimental CNV[9,10], was not affected by blockade of CCR3. Together, these experiments point to a direct effect of CCR3-induced signalling on blood vessels rather than an indirect effect through the modulation of inflammatory cells. The authors suggest that another chemokine, CXCL9 — which blocks eotaxin–CCR3 signalling in eosinophils and is expressed locally — might impair the influx of eosinophils and mast cells. They also propose that recruitment or vascular adhesion of eosinophils and mast cells could be impaired in the CNV microenvironment.

Takeda *et al.*[3] exploit their findings by developing a bioimaging tool that can detect the earliest stage of CNV, which they test in a mouse model of spontaneously developing CNV[11]. The authors tethered anti-CCR3 antibody fragments (Fabs) to tiny semiconductor nanocrystals called quantum dots (QDots). They injected these QDot–CCR3 Fabs intravenously and then performed angiography of the back of the eye. Using this tool, they could detect CNV that had not yet invaded the retina and that was invisible with the fluorescent-dye-based angiography traditionally used to visualize the process. This elegant bioimaging approach could be modified for use in humans to detect early CNV lesions before retinal damage occurs. Similar strategies might be used to identify other processes involved in the progression of AMD, such as the expression of complement proteins — activation of the complement pathway seems to switch on new blood-vessel formation in AMD[12].

It remains to be determined why CCR3

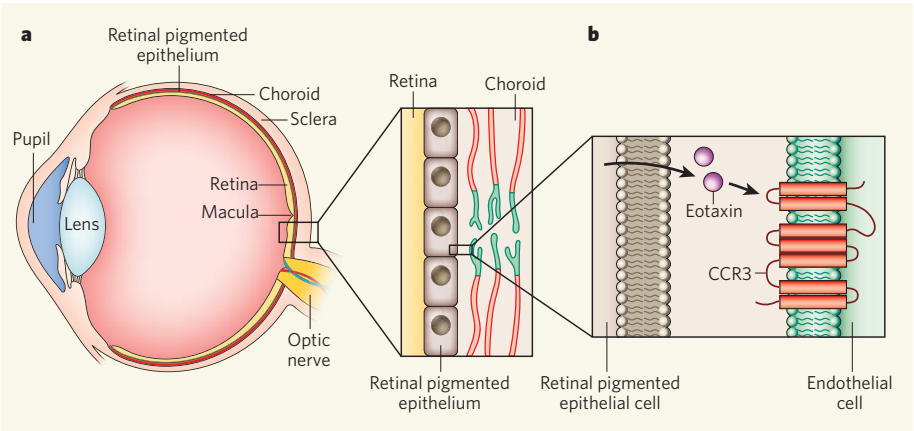


Figure 1 | CCR3 expression in macular degeneration. a, In the wet form of age-related macular degeneration, blood vessels proliferate in the choroid in a process called choroidal neovascularization (CNV). The new choroidal vessels (green) leak or bleed into the underlying retina, damaging the retina, including the central macula region. **b**, Takeda *et al.*[3] show that CCR3 (a G-protein-coupled receptor with seven transmembrane regions) is expressed on the surface of proliferating choroidal endothelial cells but not on quiescent cells. CCR3 binds to eotaxins that are secreted by the retinal pigmented epithelial cells, suggesting that the eotaxin–CCR3 axis is functional in CNV.

and its binding ligands are upregulated in CNV. The initiating event in CNV might be a physical response of the choroid to microscopic fractures in Bruch's membrane — the supporting basement membrane of the retinal pigmented epithelium — with repositioning of a single choroidal vessel that feeds new vessel formation. Could such repositioning result in endothelial expression of CCR3? Or do the endothelial progenitor cells that are incorporated into the new CNV vessels express CCR3? Evaluation of CCR3 expression on circulating endothelial progenitor cells and measurement of the amounts of circulating CCR3 ligands could address this question and possibly identify people who are at risk of AMD.

Disturbed blood-vessel formation is involved in the development of other diseases, including cancer[13], and it would be interesting to explore whether CCR3 is differentially expressed in tumour vasculature. As Takeda *et al.*[3] point out, a search for variations in the gene encoding CCR3 or in related genes in populations with AMD may also aid our understanding of what causes AMD.

The findings of Takeda and colleagues[3] suggest that the same anti-CCR3 antibodies used to detect CNV in its earliest stages could be used to treat patients with AMD before blood vessels have fully infiltrated the central macula and caused vision loss. By identifying a unique signature for CNV, the authors have triggered a paradigm shift in AMD research and have provided a wealth of questions to be answered. ■

Maria Grant is in the Department of Pharmacology and Therapeutics, University of Florida, Gainesville, Florida 32610, USA.
e-mail: grantma@ufl.edu

1. Smith, W. *et al. Ophthalmology* **108,** 697–704 (2001).
2. Rein, D. B. *et al. Arch. Ophthalmol.* **127,** 533–540 (2009).
3. Takeda, A. *et al. Nature* **460,** 225–230 (2009).
4. Rosenfeld, P. J. *et al. N. Engl. J. Med.* **355,** 1419–1431 (2006).
5. Saint-Geniez, M. *et al. PLoS One* **3,** e3554 (2008).
6. Ma, W. *et al. J. Clin. Invest.* **109,** 621–628 (2002).
7. Humbles, A. A. *et al. Proc. Natl Acad. Sci. USA* **99,** 1479–1484 (2002).
8. Harrington, P. M. *et al. Int. J. Exp. Pathol.* **80,** 177–185 (1999).
9. Zhou, J. *et al. Mol. Vis.* **11,** 414–424 (2005).
10. Sakurai, E., Anand, A., Ambati, B. K., van Rooijen, N. & Ambati, J. *Invest. Ophthalmol. Vis. Sci.* **44,** 3578–3585 (2003).
11. Ambati, J. *et al. Nature Med.* **9,** 1390–1397 (2003).
12. Nozaki, M. *et al. Proc. Natl Acad. Sci. USA* **103,** 2328–2333 (2006).
13. Carmeliet, P. *Nature* **438,** 932–936 (2005).

ORGANIC CHEMISTRY

Forgotten hydrocarbons prepared

Henning Hopf

Dendralene hydrocarbons have a reputation for being difficult — it seemed that these molecules couldn't easily be made. A practical synthesis of dendralenes opens them up for study, and reveals some surprises.

The presence of unsaturation — double or triple bonds — in molecules often largely determines their structural properties and chemical behaviour. Unsaturated hydrocarbons are well known to chemists, but one class, known as dendralenes, has been neglected, in part because the compounds were thought to be unstable. In *Angewandte Chemie*, Payne *et al.*[1] report a practical synthesis of dendralenes, and find that they are stable after all. Intriguingly, the physical and chemical properties of the compounds depend on whether there is an odd or even number of double bonds in the molecules.

There are six different ways in which carbon–carbon double bonds (C=C bonds) can be assembled to form unsaturated hydrocarbons[2]. The simplest is to connect these C=C bonds using carbon–carbon single bonds, to produce chain-like molecules known as acyclic conjugated polyolefins (Fig. 1a). Some crucial naturally occurring compounds, such as vitamin A and β-carotene, are derivatives of this class, whereas polymeric versions are familiar to materials scientists as 'organic metals' — so-called because of their conducting properties.

Alternatively, C=C bonds may be connected together using single bonds to form ring-shaped, or cyclic, molecules (Fig. 1b). The resulting hydrocarbons, known as annulenes, are either aromatic (for those with an odd number of C=C bonds), or antiaromatic (for flat molecules that have an even number of C=C bonds). Annulenes have had pivotal roles in the development of theories of the structure and reactivity of organic molecules. More practically, the aromatic hydrocarbons benzene, toluene and xylene are important feedstocks for the chemical industry.

Then there are the radialene[3] and fulvene[2] classes of hydrocarbons, which have more exotic-looking structures. In radialenes, C=C bonds 'radiate' from a central ring of carbon atoms that is formed from carbon–carbon single bonds (Fig. 1c), whereas, in fulvenes, C=C bonds radiate from annulene-like rings (Fig. 1d). Dendralenes[4] (Fig. 1e) are chain-like versions of radialenes — on paper, dendralene structures look like radialenes in which a single bond has been broken (although it isn't actually possible to convert a radialene to a dendralene). The final structural arrangement of hydrocarbons occurs when two or more C=C bonds share a common carbon atom; allenes contain two C=C bonds, whereas cumulenes contain three or more[5] (Fig. 1f).

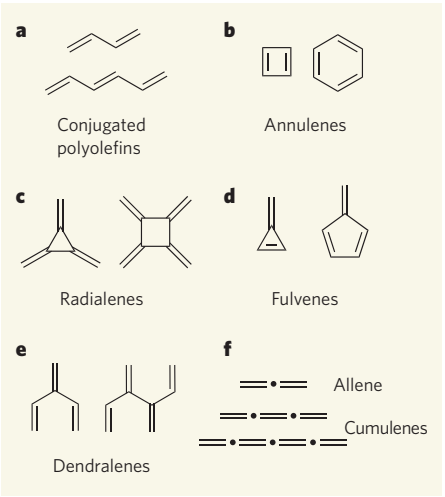


Figure 1 | Six varieties of unsaturation. **a–f,** Carbon–carbon double bonds can be assembled in six different ways to construct distinct families of unsaturated hydrocarbons, examples of which are shown. In **f**, the first member of the family is known as allene, whereas the others are known as cumulenes. Carbon atoms shared by two double bonds are represented as bold dots.

Although dendralenes exhibit a curious electronic phenomenon known as cross-conjugation — a feature also found in numerous organic dyestuffs — they have long been neglected by organic chemists. The reason is simple: the compounds could not be made readily in sufficient amounts for further study, and were assumed to be too unstable to handle under normal laboratory conditions (on the basis of what was known from the few reported examples of dendralenes[4]). Payne *et al.*[1] show that this assumption is wrong. They devised a general method for the preparation of dendralenes, and used it to make the first six members of the family. Their synthetic route provides more than enough material for further studies of the reactivity and structures of these mysterious compounds.

Payne *et al.* constructed their compounds from molecular building blocks that already contained one or more C=C bonds. For example, they used a magnesium-containing reagent (easily made from a commercially available compound) as the source of a diene fragment, which contains two C=C bonds connected by a single bond (Fig. 2a). They reacted this with other double-bond-containing compounds — various vinyl halides — in a nickel-catalysed process[6,7] that 'stitched' together the unsaturated hydrocarbon groups. In this way, Payne and colleagues prepared dendralenes containing three to five C=C bonds in good yields. The authors prepared higher oligomers (containing up to eight C=C bonds) using similar processes, providing each member of the series in gram quantities and as analytically pure substances. Previously, only milligram quantities could be made.

Like annulenes, the physical and chemical

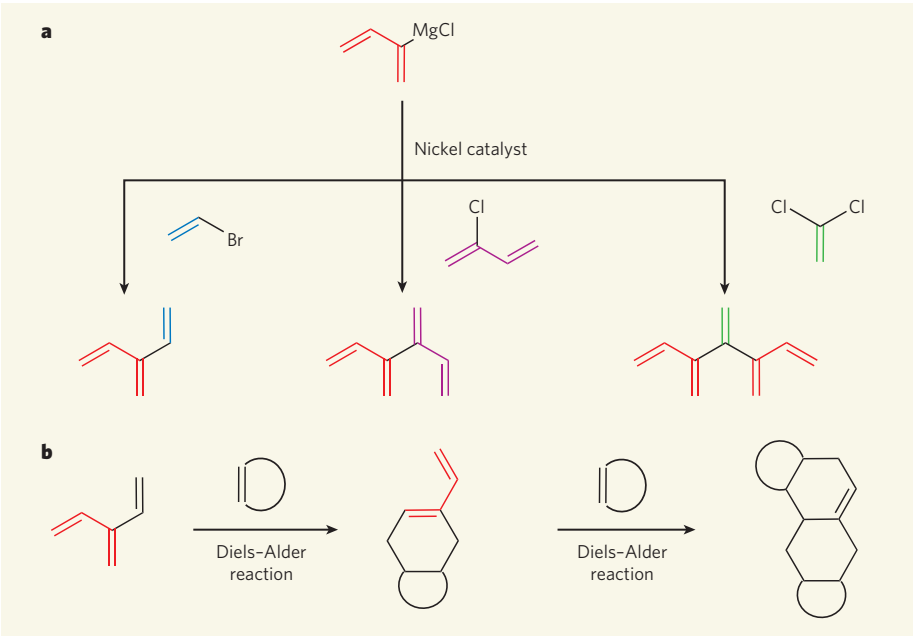


Figure 2 | Preparation and reactions of selected dendralenes. Payne *et al.*[1] have prepared dendralenes by stitching together unsaturated hydrocarbon fragments from other compounds. **a**, In these examples, the diene fragment (red) of a magnesium-containing compound is coupled in nickel-catalysed reactions to hydrocarbon fragments (various colours) of halogen-containing compounds, to make the first three members of the dendralene family. **b**, The authors also investigated the reactivities of dendralenes in Diels–Alder additions. In these reactions, a diene fragment (red) reacts to form a six-membered ring. Another diene is formed in the product, which can, in principle, take part in another Diels–Alder reaction. The cycle continues until no more dienes are formed.

properties of the newly prepared dendralenes depend on the number of C=C bonds in the molecule: the properties of the even-numbered members of the series are distinctly different from those of its odd-numbered members. A good example is the thermal stability of the compounds. Dendralenes that have an even number of C=C bonds can be kept at room temperature for weeks without any significant decomposition, whereas their odd analogues have much shorter half-lives. A similar dichotomy occurs for the electronic spectra of these compounds, and in their chemical behaviour.

Perhaps the most likely initial use of dendralenes will be in organic synthesis, acting as sources of dienes in 'cycloaddition' reactions. The most widely used cycloaddition reaction is the Diels–Alder addition, because this is the best method for preparing rings of six carbon atoms. When dendralenes are used in Diels–Alder additions, the reaction product will contain a new diene fragment, which can in principle undergo another Diels–Alder addition, and so on, until no more diene units can be generated (Fig. 2b). Such 'diene-transmissive Diels–Alder processes'[8] allow the rapid generation of molecular complexity from relatively simple starting materials in a one-pot operation.

Payne *et al.*[1] found that the reactivity of dendralenes in Diels–Alder additions again depends on the number of C=C bonds in the molecule: odd-numbered dendralenes react faster than their even-numbered counterparts. Furthermore, only the endmost dienes of odd-numbered dendralenes take part in reactions, whereas diene subunits throughout the even-numbered dendralenes react. The authors rationalized this surprising chemical effect using quantum mechanical calculations, which suggest that the geometries of the bonds in the dendralenes are at least partly responsible. In the odd-numbered dendralenes, the endmost diene subunits adopt a conformation that has long been known to be optimal for Diels–Alder reactions. These subunits therefore react quickly, and preferentially to the other diene subunits. But all of the diene subunits in the even–numbered dendralenes adopt an unfavourable conformation for Diels–Alder additions; their reactions are therefore slower than in the odd-numbered dendralenes, and no particular diene subunit reacts preferentially to the others.

With the dendralenes now available in sufficient amounts for further study, we can expect the discovery of many new reactions. The resulting products should show interesting chemical and structural properties, and would not have been available using conventional methods of synthesis. ■

Henning Hopf is at the Institute of Organic Chemistry, Technical University Braunschweig, Hagenring 30, D-38106 Braunschweig, Germany.
e-mail: h.hopf@tu-bs.de

1. Payne, A. D., Bojase, G., Paddon-Row, M. N. & Sherburn, M. S. *Angew. Chem. Int. Edn* **48,** 4836–4839 (2009).
2. Hopf, H. *Classics in Hydrocarbon Chemistry* (Wiley-VCH, 2000).
3. Maas, G. & Hopf, H. *Chemistry of Dienes and Polyenes* Vol. 1 (Ed. Rappoport, Z.) 927–977 (Wiley, 1997).
4. Hopf, H. *Angew. Chem. Int. Edn* **23,** 948–960 (1984).
5. Krause, N. & Hashmi, A. S. K. (Eds) *Modern Allene Chemistry* Vol. 1 & 2 (Wiley-VCH, 2004).
6. Corriu, R. J. P. & Masse, J. P. *J. Chem. Soc. Chem. Commun.* 144a (1972).
7. Tamao, K., Sumitani, K. & Kumada, M. *J. Am. Chem. Soc.* **94,** 4374–4376 (1972).
8. Tsuge, O., Wada, E. & Kanemasa, S. *Chem. Lett.* **12,** 1525–1528 (1983).

IMMUNOLOGY

B cells break the rules

Marilyn Diaz and Janssen Daly

A study of lymphocytes that lack a DNA-repair enzyme challenges long-standing dogma about the spatial separation of processes that rearrange antibody genes, and provides clues about the origins of B-cell cancers.

Long-lived organisms are constantly being attacked by a myriad of pathogens that have evolved mechanisms to evade the host immune system. To counter this onslaught, vertebrate T and B lymphocytes have an extraordinarily diverse repertoire of surface receptors that recognizes an array of foreign antigens. The generation of this wide range of surface B-cell receptors (membrane-bound immunoglobulin) takes place in developing B lymphocytes in the bone marrow through a process that involves breakage and recombination of variable (V), diversity (D) and joining (J) segments of immunoglobulin genes. Mature B cells in peripheral tissues (the spleen and lymph nodes) also rearrange immunoglobulin genes by DNA breakage and repair, but through a different mechanism — class-switch recombination.

In an exciting study in this issue (page 231), Wang *et al.*[1] find that a special type of V(D)J recombination — receptor editing — can take place in the periphery in mature B cells that are simultaneously undergoing class-switch recombination. In the absence of a DNA-repair enzyme, these cells experienced frequent chromosome translocations at the sites of immunoglobulin genes. These findings refute the long-standing belief that receptor editing and class-switch recombination are restricted to distinct anatomical locations and specific stages of B-cell development, and provide insight into the mechanism of gene translocations.

The immunoglobulin molecule (antibody)

consists of two heavy-chain proteins that are noncovalently bound to two light chains (either two λ- or two κ-light chains). The genes encoding the heavy chain undergo V(D)J recombination, and those encoding the light chains VJ recombination, to form the V(D)J exon, which encodes the region of the immunoglobulin molecule that determines its specificity. This reaction is initiated by the RAG enzyme complex, which induces double-stranded DNA breaks in the V, D and J regions (Fig. 1a). After recombination, the breaks are repaired through a process known as non-homologous end-joining (NHEJ). In receptor editing, developing B cells in the bone marrow undergo successive rounds of RAG-mediated V(D)J recombination to exchange the light chains of an autoreactive immunoglobulin molecule so that it is no longer autoreactive[2,3].

Class-switch recombination, which changes the effector function of immunoglobulins, is a process by which the exons for the constant domain of the heavy chain of IgM are swapped with downstream exons to generate different classes of antibody, such as IgG, IgE or IgA. Class-switch recombination is also initiated by DNA breaks, in this case induced by the AID enzyme (Fig. 1b). These breaks can also be repaired by NHEJ or by an alternative end-joining (A-EJ) pathway. Another AID-mediated mechanism of gene rearrangement in peripheral B cells is somatic hypermutation. Here, mutations accumulate in the rearranged immunoglobulin genes, potentially increasing antibody-binding specificity.

That immunoglobulin-gene rearrangements are associated with double-stranded DNA breaks underscores the enormous selective pressures driving the evolution of these processes — when not repaired correctly, DNA breaks can lead to chromosome translocations, which predispose to cancer. Indeed, certain types of human B-cell tumour (lymphomas) frequently contain translocations that merge antigen-receptor genes with a proto-oncogene (a gene with the potential to promote cancer).

There are several biological mechanisms that reduce the tumour-causing potential of DNA breaks in B cells, including restricting these processes to distinct tissues, such as the bone marrow for V(D)J recombination and peripheral tissues for class-switch recombination. However, as Wang *et al.*[1] show, these apparent safeguards underestimate the plasticity of B cells.

The authors examined mice in which the NHEJ double-strand-break-repair protein XRCC4 is deleted in mature B cells. They report that a subset of activated peripheral B cells with defective NHEJ simultaneously harbour double-stranded DNA breaks associated with V(D)J recombination and class-switch recombination. Surprisingly, when these cells are activated by signals that lead to class-switch recombination in the DNA locus encoding the immunoglobulin heavy chain (*Igh*), they also re-initiate V(D)J recombination at the immunoglobulin-λ light-chain locus (*Igl*).

In Wang and colleagues' study, the splenic B cells that reactivate V(D)J recombination are not undergoing the conventional V(D)J recombination used by developing B cells to generate the initial immunoglobulin repertoire. Instead, the authors argue, these cells are undergoing receptor editing, which was thought to be confined to immature B cells in the bone marrow. It has previously been suggested[4] that editing can occur in peripheral B cells during the generation of memory B cells in specific regions of the spleen — the germinal centres — through a mechanism termed receptor revision. However, the peripheral editing in the B cells in Wang *et al.*'s study[1] seems to be distinct from the receptor-revision mechanism, because the B cells lack germinal-centre markers and are not activated by signals that normally lead to germinal-centre formation.

Incorrect repair of the breaks initiated by receptor editing and class switching frequently resulted in chromosome translocations[1] involving *Igh* and *Igl* (Fig. 2, overleaf). Although neither *Igh* nor *Igl* are chromosomal regions that promote cancer, translocations involving *Igh* or *Igl* with a proto-oncogene, such as *c-myc*, can result in lymphomas. Indeed, deletion of both *Xrcc4* and *p53* (a gene encoding a tumour-suppressor protein) in mature B cells in mice leads to lymphomas, known as CXP lymphomas, the cells of which reveal evidence of receptor editing and class switching[5]. Thus it is likely that the XRCC4-depleted cells in Wang and colleagues' study are the progenitors of mouse CXP lymphoma cells, and that B cells with similar mutations may contribute to some human B-cell lymphomas.

B cells that undergo both class-switch recombination and receptor editing harbour AID- and RAG-dependent DNA breaks, and Wang *et al.*[1] capitalize on this characteristic to identify factors that may enhance translocations between immunoglobulin loci. One such factor may be the A-EJ pathway of DNA double-strand-break repair. As the XRCC4-depleted cells cannot repair DNA breaks with NHEJ and are forced to use the A-EJ pathway, the increase in translocations may reflect a propensity of the A-EJ mechanism to generate such translocations.

The authors found a strong correlation between translocations and proximity of *Igh* and *Igl* in the nucleus of B cells at interphase, a cell-cycle stage during which much of the gene expression occurs. Similarly, co-localization of *c-myc* and *Igh* in the nucleus correlated with translocations between these loci, although the rate-limiting factor was the frequency of breaks at *c-myc*, which is strongly AID-dependent[6].

That the B cells studied by Wang *et al.*[1] are

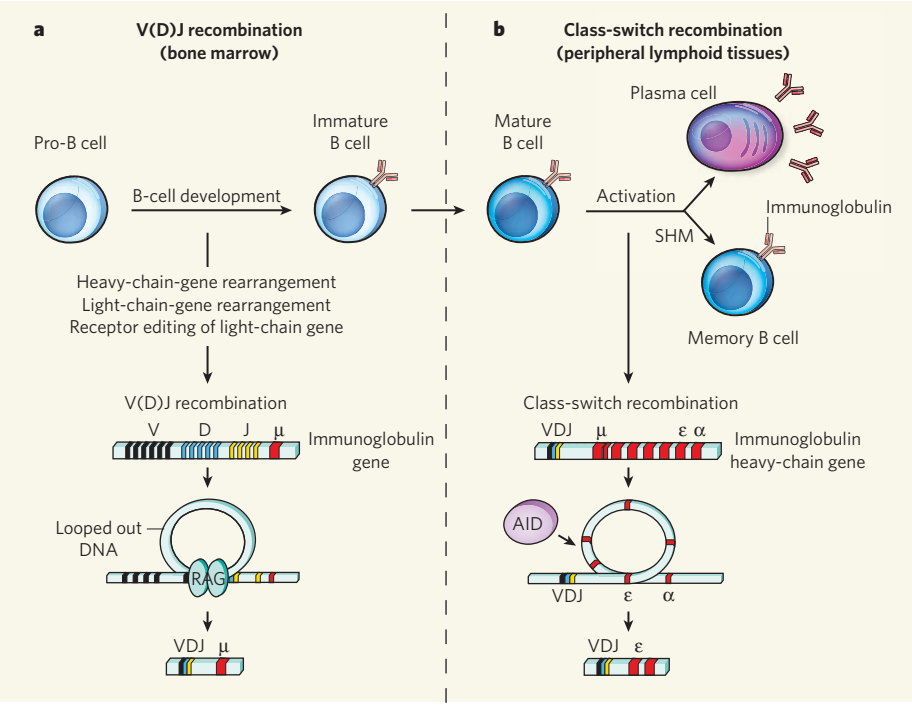


Figure 1 | Immunoglobulin gene rearrangements. a, Recombination of variable (V), diversity (D) and joining (J) segments of immunoglobulin genes generate B-cell receptors during development in the bone marrow. If the immunoglobulin (antibody) on the developing B cell reacts against 'self' antigen, the cell undergoes further light-chain-gene recombination (receptor editing), to generate a non-autoreactive immunoglobulin. **b,** Class-switch recombination occurs after activation of mature B cells in peripheral lymphoid tissues (the spleen and lymph nodes). In class switching, the μ exons are swapped with downstream exons to generate a different antibody class. Activated B cells also undergo somatic hypermutation (SHM) as they develop into memory B cells. Whereas the RAG proteins initiate V(D)J recombination (**a**), class-switch recombination and somatic hypermutation are triggered by the AID enzyme (**b**). All processes are initiated by or involve DNA double-strand breaks. Wang *et al.*[1] find that, contrary to long-held dogma, mature B cells can simultaneously undergo both class switching and receptor editing in the periphery.

NATURE|Vol 460|9 July 2009

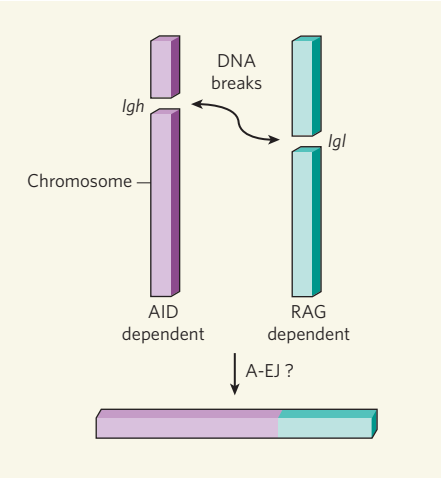


Figure 2 | Chromosome translocations in B cells. Wang *et al.*[1] show that mature B cells that lack an essential DNA-repair enzyme undergo both AID-induced class-switch recombination at the immunoglobulin heavy chain locus (*Igh*) and RAG-induced receptor editing at the immunoglobulin-λ light-chain locus (*Igl*). These breaks can lead to chromosome translocations involving *Igh* and *Igl*. Use of the alternative end-joining (A-EJ) pathway may contribute to the formation of such translocations.

the progenitors of mouse CXP lymphomas suggests that peripheral editing occurs *in vivo* and may contribute to the development of such cancers. Paradoxically, in the authors' study, the signals activating V(D)J recombination in B cells are typically associated with activation of class-switch recombination rather than autoreactivity — the trigger for receptor editing in the bone marrow. So, if not the revision of an autoreactive receptor, what is achieved by replacing the light chain in this subset of peripheral B cells? To fully understand the significance of this phenomenon, it will be important to determine the frequency of receptor editing in peripheral B cells and its physiological function. Finally, this novel B-cell population can be exploited to elucidate the mechanisms that promote translocations between antigen-receptor loci and proto-oncogenes. ■

Marilyn Diaz and Janssen Daly are at the Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences of the National Institutes of Health, Research Triangle Park, North Carolina 27709, USA.
e-mails: diaz@niehs.nih.gov; dalyj2@niehs.nih.gov

1. Wang, J. H. *et al. Nature* **460,** 231–236 (2009).
2. Gay, D., Saunders, T., Camper, S. & Weigert, M. *J. Exp. Med.* **177,** 999–1008 (1993).
3. Tiegs, S. L., Russell, D. M. & Nemazee, D. *J. Exp. Med.* **177,** 1009–1020 (1993).
4. Hertz, M. & Nemazee, D. *Curr. Opin. Immunol.* **10,** 208–213 (1998).
5. Wang, J. H. *et al. J. Exp. Med.* **205,** 3079–3090 (2008).
6. Robbiani, D. F. *et al. Cell* **135,** 1028–1038 (2008).

MATERIALS SCIENCE

Nanotubes sorted using DNA

Mark C. Hersam

A vast number of DNA sequences are possible, and so finding the few that bind to a particular non-DNA entity is a daunting task. A systematic search algorithm has found sequences that target specific carbon nanotubes.

For nearly two decades, the carbon nanotube has been the poster child of nanotechnology. Researchers have used its exemplary physical and chemical properties in a diverse range of prototype devices, spanning such technologies as alternative energy, biotechnology and computing. Underlying this success is the exquisite sensitivity of the nanotubes' properties to their physical size and atomic structure. However, this sensitivity also creates a fundamental problem: because current syntheses of carbon nanotubes lack atomic-level control, samples produced are mixtures of nanotubes of different sizes and atomic geometries, and thus possess non-uniform properties. This non-uniformity has confounded their use in large-scale commercial applications, which invariably require materials that have consistent, reproducible performance.

Many researchers have therefore devised schemes for sorting carbon nanotubes according to their physical and electronic structures[1]. Inspiration has often come from bioseparation methods, leading to the use of electrophoresis[2], ultracentrifugation[3] and chromatography[4] techniques. DNA has had a recurring supporting role in these studies because of its ability to disperse carbon nanotubes in biologically compatible aqueous solutions[5–7]. But despite its ability to bind to specific molecules depending on its base sequence, DNA has not been systematically explored as a means of isolating different types of carbon nanotube — until now. On page 250 of this issue, Tu *et al.*[8] describe the heroic efforts that resulted in their identifying more than 20 DNA sequences that each selectively bind a specific carbon-nanotube structure. Their careful study uncovers distinct patterns of DNA sequences that will inform future efforts in nanotube separation, and provides fundamental insight into the chemical interactions between arguably the most important biomolecule and one of the most-studied nanomaterials.

To appreciate the magnitude of the authors' task, consider that custom-made DNA sequences containing 100 nucleotides are readily available commercially. Because there are four DNA bases — adenine (A), thymine (T), guanine (G) and cytosine (C) — this amounts to 4^{100} (10^{60}) sequences that could be screened for their nanotube-binding properties. This number is almost unfathomably large, and so the authors had to devise a systematic method to focus their search before they could attack this problem experimentally.

Initially, Tu *et al.* limited their search to DNA molecules containing 28 or 30 bases, thus restricting the number of possibilities to 4^{30} (10^{18}). Although this is a huge improvement over 10^{60}, further refinement was clearly necessary. The authors therefore used a sequence-pattern-expansion scheme to come up with a manageable set of DNA sequences, starting with simple patterns and then adding complexity in a confined, progressive way. The scheme started with molecules that contained only one kind of base, thus yielding four sequences. Complexity was added in the next phase of the scheme — the second-order expansion — when all 16 variants of two-base repeats were added (for example, $(AT)_{15}$). By following this procedure to a third and fourth order of complexity, Tu *et al.* constructed a search set containing approximately 350 different DNA sequences.

The authors used each of these sequences to disperse a randomly produced mixture of carbon nanotubes in water. They then used chromatography to separate the resulting 350 solutions into fractions based on the ionic charge of the solutes, and characterized each fraction spectroscopically to see if any of the DNA sequences had formed complexes specifically with a single kind of nanotube. Although most of the sequences had not, a series of DNA molecules that contained alternating patterns of one or more purines (A or G) and pyrimidines (T or C) — such as $(GT)_{15}$, $(TCG)_{10}$ and $(ATTT)_7$ — showed a differential affinity for nanotubes as a function of nanotube structure.

Recognizing the successful purine–pyrimidine motifs, Tu *et al.* performed more experiments in which they varied the length of their DNA sequences, and found that shorter DNA molecules (as short as eight bases) bind to nanotubes with exceptional selectivity. In all, more than 20 distinct DNA sequences selected one kind of carbon nanotube from an as-prepared mixture. The purity of semiconducting nanotubes isolated in this way approached 99%, equalling or exceeding those obtained by all previous carbon-nanotube sorting techniques[1].

Although the molecular-recognition mechanism involved in this DNA–nanotube binding[8] is not fully understood, highly suggestive trends can be identified from the successful DNA sequences. For example, DNA molecules that contain alternating purine–pyrimidine patterns form stable, well-ordered, two-dimensional sheets through hydrogen bonding (see Fig. 2a on page 252) — structures that resemble the ubiquitous β-sheet motif in proteins. Fur-

thermore, these DNA sheets are expected to form stable cylindrical structures reminiscent of the barrel-shaped structures formed from β-sheets in proteins. Such DNA barrels could thus encapsulate cylindrical carbon nanotubes, presumably with high affinity for nanotubes that have diameters that match the inner diameter of the barrel (see Fig. 2c on page 252). This structural mechanism is different from those of previously described methods for separating carbon nanotubes (which are based on differential chemical binding affinity[1]) and thus might explain the exceptional purities achieved by Tu and colleagues.

The current study[8] marks a considerable advance in the carbon-nanotube field, but major issues remain unresolved. For example, carbon nanotubes are chiral[9,10] — each type of nanotube exists as one of two mirror-image forms depending on the direction in which its carbon atoms coil up to form the tube. So far, DNA has not been shown to be able to distinguish between the mirror-image forms of nanotubes, which means that the DNA-separated nanotubes might be sub-optimal for some optical-device applications.

In addition, the DNA sequences identified by Tu *et al.*[8] show higher selectivity for semiconducting carbon nanotubes than for those that have metal-like conductivities, and so further work is required to isolate both types equally using DNA. The authors' approach is also relatively expensive (because of the high cost of DNA), which might limit its use in large-scale applications. The ultimate solution to sorting carbon nanotubes will therefore probably be a hybrid method that combines the best attributes of several different techniques[1]. In the meantime, Tu and colleagues' approach possesses clear advantages for single-step, low-quantity separations that will be of great interest to research groups around the world. ■

Mark C. Hersam is in the Departments of Materials Science and Engineering, and of Chemistry, Northwestern University, Evanston, Illinois 60208-3108, USA.
e-mail: m-hersam@northwestern.edu

1. Hersam, M. C. *Nature Nanotechnol.* **3,** 387–394 (2008).
2. Tanaka, T., Jin, H., Miyata, Y. & Kataura, H. *Appl. Phys. Express* **1,** 114001 (2008).
3. Arnold, M. S., Green, A. A., Hulvat, J. F., Stupp, S. I. & Hersam, M. C. *Nature Nanotechnol.* **1,** 60–65 (2006).
4. Zheng, M. & Semke, E. D. *J. Am. Chem. Soc.* **129,** 6084–6085 (2007).
5. Zheng, M. *et al. Nature Mater.* **2,** 338–342 (2003).
6. Zheng, M. *et al. Science* **302,** 1545–1548 (2003).
7. Arnold, M. S., Stupp, S. I. & Hersam, M. C. *Nano Lett.* **5,** 713–718 (2005).
8. Tu, X., Manohar, S., Jagota, A. & Zheng, M. *Nature* **460,** 250–253 (2009).
9. Peng, X. *et al. Nature Nanotechnol.* **2,** 361–365 (2007).
10. Green, A. A., Duch, M. C. & Hersam, M. C. *Nano Res.* **2,** 69–77 (2009).

QUANTUM INFORMATION

Circuits that process with magic

Raymond W. Simmonds and Frederick W. Strauch

Practical quantum computation will require a scalable, robust system to generate and process information with precise control. This is now possible using a superconducting circuit and a little quantum magic.

Have you ever wondered what quantum computers will really look like? Will they be warehouses full of vacuum chambers, lasers and optics tables? What will be doing the computing? Thus far, simple quantum algorithms have been performed with small numbers of molecules[1], cold ions[2] or photons[3] — systems that were designed by nature but that can be cleverly controlled. Scaling these systems up to a workable quantum computer will require a little error correction and a lot of hard work.

This challenge may have become a little easier. Superconducting circuits[4] offer a unique platform for constructing custom-designed, fully engineered, scalable quantum systems. These systems involve electrical circuits patterned on a microchip, similar to the classical processors of today. The circuit used by DiCarlo *et al.*[5], described on page 240 of this issue, which incorporates two quantum bits (qubits) on either side of an extended, resonant microwave cavity, can be controlled by tabletop electronics and has allowed these authors to demonstrate the first superconducting two-qubit quantum-information processor.

The processing of quantum information involves 'magic tricks' that are only possible through careful control of a quantum system[6]. Finding a system that maintains quantum coherence, the magic spell, over relatively long periods of time — long enough to use the magician's tricks — is a challenge. DiCarlo and colleagues[5] have developed 'transmon' qubits, tiny superconducting circuits that routinely hold on to quantum coherence for a microsecond[7]. Although this is shorter than the millisecond (or longer) timescales obtained in atomic and photonic systems, superconducting circuits benefit from nanosecond timescales for single-qubit control and pairwise qubit interactions provided by fast, commercial electronics. These control times are orders of magnitude faster than those for most atomic systems.

Single qubits have two 'basis states', a ground state $|0\rangle$ and an excited state $|1\rangle$. A two-qubit system is described with four basis states $|0,0\rangle$, $|0,1\rangle$, $|1,0\rangle$ and $|1,1\rangle$, as denoted by $|$left qubit, right qubit$\rangle$. In quantum mechanics, an arbitrary two-qubit state is formed with combinations of these basis states: $a|0,0\rangle + b|0,1\rangle + c|1,0\rangle + d|1,1\rangle$, where the four coefficients a, b, c and d are complex numbers with a particular amplitude and phase. The computational elements (gates) of quantum algorithms are divided into single-qubit rotations (altering single-qubit states, such as flipping a bit $|0\rangle \leftrightarrow |1\rangle$) and two-qubit operations (altering two-qubit states).

Previously, the authors have shown high-fidelity benchmarking of single-qubit rotations[8]. In their new study[5], DiCarlo *et al.* implement a two-qubit operation, a trick known as a controlled phase (C-Phase) gate. By tuning a magnetic flux on each qubit, the energy of the two-qubit states can be controlled directly. Here, the presence of the cavity provides a two-qubit interaction, helping to uniquely determine these energy values. Because quantum states evolve in time (t) according to $e^{-iEt/\hbar}$, where E is the energy of the state, i is the imaginary unit and $\hbar$ is Planck's constant, this leads to a phase change in the coefficient of each two-qubit basis state. Careful control and timing of these flux excursions allows a specific two-qubit basis state to be targeted for an overall π phase change or complete sign reversal ($\pm$ to $\mp$). Measurement of the two-qubit states is achieved by passing microwave photons through the resonant cavity containing the two qubits[9,10].

Consider the blackjack-style card game analogue shown in Fig. 1. Imagine a deck of many 'quantum cards', where each card has a colour (red or blue), a rank (0 or 1) and a suit (+ or –). A two-qubit quantum state can be represented by a set of up to four possible two-card plays, in which the two colours represent the left and right qubit respectively, the rank represents the individual qubit states, and the suit indicates a phase factor. A 'ground-state' hand — a single two-card play {(+0,+0)} (corresponding to $|0,0\rangle$) — occupies one square of the card table (Fig. 1a). In this game, the player can ask the dealer for a 'hit' (a single-qubit rotation) on a chosen colour (qubit), in which case additional cards are added to the table. Each square with a ±0-card (±1-card) adds a ±1-card (∓0-card) paired with a copy of the other colour card from that square. These additional two-card plays are then distributed to the appropriate squares on the card table. The hand can then be further refined, removing cards if necessary (see Fig. 1 for rules).

Hits on both colours, starting from the ground-state hand, lead to four single-qubit cards {([+0,+1],[+0,+1])} (Fig. 1b). Once paired with each other and distributed, four two-card plays {(+0,+0),(+0,+1),(+1,+0),(+1,+1)} occupy each square of the card table (Fig. 1c). Note that this hand corresponds to the two-qubit 'product state' $|0,0\rangle + |0,1\rangle + |1,0\rangle + |1,1\rangle$ and has an

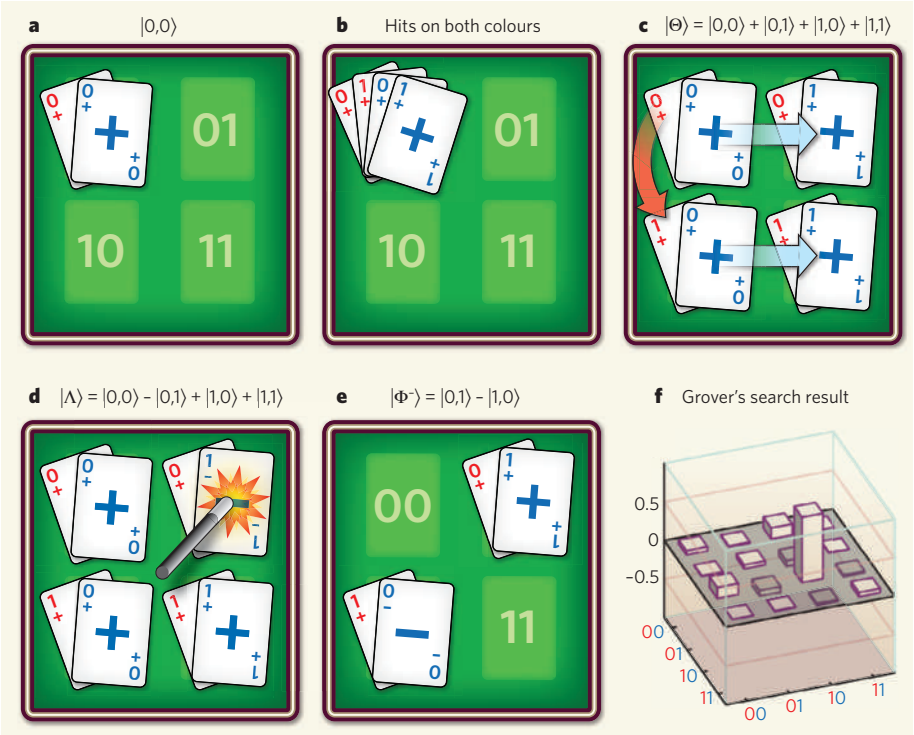


Figure 1 | A blackjack-style card game analogue for a two-qubit quantum processor. A two-qubit quantum state can be represented by pairs of 'quantum cards', where each card has a colour (red or blue), a rank (0 or 1) and a suit (+ or −), corresponding respectively to a left or right qubit, a single qubit state, and a phase factor. A 'hand' occupies up to four possible squares on the card table for each possible two-card play. When a player asks for a 'hit' on a chosen colour, equivalent to a single-qubit rotation, cards are added to the table. Each square with a ±0-card (±1-card) adds a ±1-card (∓0-card) paired with a copy of the other colour card from that square. These new two-card plays are then distributed to the appropriate squares on the card table (distributive rule). After a hit, multiple two-card plays in each square are refined. Any single two-card play can have both suits changed (associative rule). One suit can be changed for each and every two-card play (overall-phase rule). Duplicate two-card plays are removed from the table (normalization rule). Pairs of two-card plays that differ by one suit are removed (addition rule). The remaining two-card plays on the table represent the hand (or final two-qubit state). **a**, The initial 'ground-state' hand {(+0,+0)}, which corresponds to the two-qubit state $|\text{left qubit, right qubit}\rangle = |0,0\rangle$. **b**, Simultaneous hits (rotations) on both colours (qubits) produce {([+0,+1],[+0,+1])}, which will be paired with each other and distributed to produce the 'product state' $|\Theta\rangle$ shown in **c**. **c**, Taken sequentially, a hit on the red cards (left qubit) of the ground-state hand produces one additional two-card play (red arrow). A further hit on the blue cards (right qubit) produces two additional two-card plays (blue arrows), leading to an equal-superposition hand (or $|\Theta\rangle = |0,0\rangle + |0,1\rangle + |1,0\rangle + |1,1\rangle$). **d**, A 'magic wand' is tapped on the two-card play (+0,+1) to produce (+0,−1), leading to an entangled hand (or a C-Phase gate targeting $|0,1\rangle$ is applied to $|\Theta\rangle$, producing $|\Lambda\rangle = |0,0\rangle - |0,1\rangle + |1,0\rangle + |1,1\rangle$). **e**, A further hit on the red cards leads to a Bell-state hand (or rotating the left qubit of $|\Lambda\rangle$ leads to $|\Phi^-\rangle = |0,1\rangle - |1,0\rangle$). **f**, DiCarlo and colleagues[5] performed a quantum algorithm called Grover's search. Beginning from $|\Theta\rangle$, a phase reversal is hidden with $|1,0\rangle$. After the algorithm, the final state tomography shows that 85% of the time, the phase reversal was found correctly.

equal collection or superposition of all ranks of a single suit. Like the distributive law in mathematics, this is a specific example of a product of single qubit states that is factorable.

A magician can now perform a quantum trick using the C-Phase gate like a magic wand. For a chosen two-card play, say (+0,+1), the magician can change the suit of a single card in the two-card play. For the equal-superposition hand described above, this trick produces {(+0,+0),(+0,−1),(+1,+0),(+1,+1)} (or the state $|0,0\rangle - |0,1\rangle + |1,0\rangle + |1,1\rangle$) (Fig. 1d). A remarkable thing has happened! The cards cannot be factored into a product state as before, because they no longer share the same suit. With a simple touch of the wand, the hand has become 'entangled'. Entanglement can be thought of as a magical interdependence between the two qubits, where the state of one qubit is linked to the other. Now, continued hits on either qubit will influence two-card plays along the diagonal squares of the card table. For example, a further hit (rotation) on the red cards (left qubit) will remove the (+0,+0) and (+1,+1) two-card plays (see Fig. 1 for rules), leading to a highly entangled hand known as a Bell state (Fig. 1e).

In their experiments, DiCarlo and colleagues[5] generated four Bell states by performing a C-Phase gate on each of the four possible two-qubit basis states. They then performed a large number of measurements to obtain tomographic pictures of the two-qubit quantum states[10,11]. The data show high-quality, clearly entangled states with fidelities of more than 90%. Next, with their C-Phase gate (magic wand) perfected, DiCarlo *et al.* decided to play a game of hide-and-seek, performing a quantum algorithm known as Grover's search[1].

Imagine our card game again. We begin with the equal-superposition hand, {(+0,+0),(+0,+1),(+1,+0),(+1,+1)}. With his eyes closed, the magician (Grover) asks the audience to tap the wand on any two-card play in the hand. The audience chooses (+1,+0), hiding a secret suit change. Grover, with eyes still closed, first hits on both colours, then taps his wand on (+0,+0) followed by another hit on both colours. This leaves only one two-card play on the table, {(+1,+0)}. Grover asks the audience: "Was this the pair of cards you tapped with the wand?" And indeed, in the experiments, as the final-state tomography shows (Fig. 1f), this was the case 85% of the time. In fact, DiCarlo and colleagues go so far as to reveal their sleight of hand entirely in their paper[5], by performing state tomography after each processing step. The somewhat more mathematical Deutsch–Jozsa algorithm[6] was also performed using the same basic magic tricks.

DiCarlo and colleagues' demonstrations provide real-world proof that quantum mechanics can be harnessed to process information. However, to use this magic for a real-world problem will require longer algorithms using larger processors (imagine, the size of the card table will double with each additional qubit — only a magician can deal with that!). The ability to perform longer algorithms is limited by a loss of coherence, largely due to microscopic defects found inside fabricated qubits[12], the sources of which are still under investigation. With many groups of researchers working towards improving superconducting circuits, we anticipate larger processors in the near future. These can be formed by wiring up more qubits to a single cavity[13], or one can imagine more complex networks of cavities[14]. The two-qubit processor of today may indeed be closer than we think to the many-qubit processor of tomorrow. ■

Raymond W. Simmonds is at the National Institute of Standards and Technology, Boulder, Colorado 80305, USA. Frederick W. Strauch is in the Department of Physics, Williams College, Williamstown, Massachusetts 01267, USA.
e-mails: raymond.simmonds@nist.gov; frederick.w.strauch@williams.edu

1. Chuang, I. L., Vandersypen, L. M. K., Zhou, X., Leung, D. W. & Lloyd, S. *Nature* **393,** 143–146 (1998).
2. Gulde, S. *et al. Nature* **421,** 48–50 (2003).
3. Kwiat, P. G., Mitchell, J. R., Schwindt, P. D. D. & White, A. G. *J. Mod. Opt.* **47,** 257–266 (2000).
4. Clarke, J. & Wilhelm, F. K. *Nature* **453,** 1031–1042 (2008).
5. DiCarlo, L. *et al. Nature* **460,** 240–244 (2009).
6. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, 2000).
7. Koch, J. *et al. Phys. Rev. A* **76,** 042319 (2007).
8. Chow, J. M. *et al. Phys. Rev. Lett.* **102,** 090502 (2009).
9. Wallraff, A. *et al. Nature* **431,** 162–167 (2004).
10. Filipp, S. *et al. Phys. Rev. Lett.* **102,** 200402 (2009).
11. Steffen, M. *et al. Science* **313,** 1423–1425 (2006).
12. Martinis, J. M. *et al. Phys. Rev. Lett.* **95,** 210503 (2005).
13. Fink, J. M. *et al.* Preprint at http://lanl.arxiv.org/abs/0812.2651 (2008).
14. Helmer, F. *et al. Europhys. Lett.* **85,** 50007 (2009).

OBITUARY

Herbert Frank York (1921–2009)

A voice of calm in the era of nuclear weapons.

Herb York, who died on 19 May at the age of 87, was present at the creation of the US postwar defence research and development establishment and, from a very early date, was one of its scientific leaders. He also quickly recognized the limitations of technology in providing security, and became a leader in arms control and in educating younger people on the complexities of our nuclear age. York was unflappable, always accessible and always open to a good argument. His initiatives in defence, arms control and education continue to shape key institutions in the United States.

During the Second World War, York worked on the Manhattan Project to develop nuclear weapons. After the atom bomb was dropped over Hiroshima on 6 August 1945, he concluded that war had finally become too terrible to fight, "not know[ing]", as he later wrote, "that the same thing had been said many times before". A sense of responsibility for the consequences of the atom bomb never left him.

In 1952, he was picked by Ernest Lawrence from the obscurity of an assistant professorship in physics at the University of California (UC), Berkeley, to lead what became the Lawrence Livermore National Laboratory. During York's six years there (Lawrence suggested he call himself director after the first year and a half), the laboratory grew from about 100 people to 3,000, and among other tasks was entrusted with the development of the warhead for the first submarine-launched missile, Polaris. The laboratory pioneered several techniques, among them the use of the largest computers then available (which were less powerful than today's laptops) to model complex events such as nuclear explosions. As with most good labs, constant review and questioning of assumptions, goals and methods were the norm. But York added a good-humoured, positive yet realistic attitude, in which setbacks were cause for learning rather than finger-pointing criticism.

In 1957, York was asked by the first presidential science adviser, James Killian, to join the President's Science Advisory Committee, which President Dwight Eisenhower had just founded in response to the Soviet launch of the Sputnik satellite. Several of the main lines of today's defence establishment were created then, particularly as regards nuclear forces and research and development. On the committee, York was a major influence in starting what is now called

BETTMANN/CORBIS

the Defense Advanced Research Projects Agency, a (comparatively) freewheeling research and development establishment within the Pentagon, which was supposed to — and to this day does — introduce new technological ideas into the system. He was promptly named as its first chief scientist and then became the first director of Defense Research and Engineering, a post from which he both supervised and moderated the burgeoning US missile forces.

York was also instrumental in transferring the US Army's space facility at Huntsville, Alabama (where the German rocket expert Wernher von Braun and his colleagues had landed at the end of the Second World War), and the Jet Propulsion Laboratory at the California Institute of Technology in Pasadena, to NASA. He thereby helped to place space exploration in the United States firmly in civilian hands.

It was from those years that York came to have a sober view of what the military–industrial–congressional complex would do without (and sometimes despite) politically powerful and technically well-informed supervision. In this he was very much in tune with Eisenhower. Both were ahead of the political establishment of the time and had mixed success.

After recovering from a serious heart attack in 1960, York became the founding chancellor of UC San Diego, a campus dedicated to excellence in science. He served four years in that post and remained at the university as professor of physics for the rest of his life. He served a second term as interim chancellor from 1970 to 1972.

Throughout that period, York continued to work on the issues raised by the nuclear age, particularly arms control and non-proliferation. Joseph Nye, then deputy to the undersecretary of state for security assistance, science and technology, and chair of the National Security Council Group on Nonproliferation of Nuclear Weapons, remembers him "as a mature, insightful and calming presence as an adviser as we developed [President] Carter's non-proliferation policy". Carter also appointed York as the ambassador leading the US delegation in negotiating a Comprehensive Test Ban Treaty in 1979.

In the early1980s, York was instrumental in getting Jerry Brown, governor of California, and UC president David Saxon to set up the Institute on Global Conflict and Cooperation (IGCC), which is dedicated to research and teaching in the area of preventing war. The institute's creation was the realization of York's goal that the Second World War, and especially Hiroshima, had inspired forty years before. It spans several UC campuses as well as the nuclear-weapons laboratories, which were then managed by the university. York served as the first director and continued to contribute to it until his death.

Several of his initiatives at the IGCC continue, such as the summer nuclear-weapons policy training programmes and the Track 2 dialogues, which are informal conversations among officials from different countries on particular issues. He also made a more personal contribution: he linked today's students and young researchers to the time when nuclear weapons were first introduced, putting a human face on how the new dilemmas were resolved or not, and why policy responses took the shape that they did. York combined realistic assessments of the dangers with advocacy for steps to alleviate them, all of this serious work being enlightened by his sunny, open disposition.

Herb York won most of the honours available in his various fields, including the 1994 Leo Szilard Lectureship Award and the 2000 Enrico Fermi Award, and published six books about his experiences and views. He exemplified how scientists can best serve the common cause in a dangerous and controversial area, despite the many puzzles and contradictions that such service implies. He did it with a sense of balance, good humour and decency under pressure that continues to encourage the rest of us.

Michael M. May
Michael M. May is at the Center for International Security and Cooperation, Stanford University, Stanford, California 94305-6165, USA.
e-mail: mmay@stanford.edu

HORIZONS

The possibility of impossible cultures

Marc D. Hauser

Insights from evolutionary developmental biology and the mind sciences could change our understanding of the human capacity to think and the ways in which the human mind constrains cultural expressions.

The mental abilities of humans and other animals seem to have far greater similarities than differences[1–5]. In the past 20 years or so, there have been reports of animals acquiring language, producing music, feeling empathy and teaching, leading to the conclusion that the differences between human and animal thought are just matters of degree. Such a continuum also seems to apply to cultural forms, such as language, music and morality. Humans generate an extraordinary range of cultural expressions and seem to have an almost unbounded capacity to do so[6–9]. Just think of the differences between musical styles — for example between baroque, bhangra, gamelan and hip hop — the variation seems to be limitless.

The idea that such variation is continuous implies that there are no meaningful attributes common to all mental abilities or all cultural forms, as well as no limits to the abilities or forms that are possible. From this viewpoint, there are no gaps in the distribution of mental capacities across species or in the distribution of potential cultural forms.

But if these commonly held ideas are approached from another perspective, the opposite conclusions can be drawn. On the basis of recent developments in evolutionary developmental biology and the mind sciences, especially linguistics, I propose that there are two crucial gaps within the range of variation: one representing psychological discontinuity between humans and all other animals, and the other representing cultural discontinuities within the range of possible cultural forms. Contemplating the possibility of cultural discontinuities forces a further consideration: some cultural forms will never be considered or, if they are, will prove problematic to acquire and sustain — these can be thought of as impossible cultures. These ideas set the stage for new approaches to understanding human thought — in terms of its distinctive characteristics and the limits it imposes on cultural expression — from the levels of genes and neurons to thoughts and behaviour.

Impossible morphologies

Nature provides a bewildering and seemingly unbounded variety of animal forms, from the microscopic (such as insects) to the macroscopic (such as dinosaurs), and from the pointy and spherical (blowfish) to the smooth and cylindrical (snakes). Until recently, the dominant idea was that variation emerged from random processes, with adaptations sculpted by the blind process of natural selection.

New molecular approaches have now sharpened our understanding of the sources of variation and of how developmental programs interact with and constrain evolutionary processes, leading to a restricted range of adaptations. Much of this work was inspired by the rich description of the Cambrian period (about 500 million years ago), in which there was a rapid and unprecedented explosion of new life forms. The fact that such variation appeared within a short time span, and that simple organisms such as worms and insects were equipped with genomes almost as large as that of humans, leads to two conclusions and raises one substantive challenge.

First, given the rapid emergence of different life forms, the cellular machinery that evolved before and during the Cambrian was highly generative. That is, it provided a massive suite of options for organisms that were confronting different ecological circumstances and challenges.

Second, although the genes encoding this machinery were in place, they were often hidden from view, their presence not evident at the level of anatomy or behaviour. This fact highlights the importance of documenting not only which regions of a genome are expressed but also which regions can, but might not, be expressed.

Although the diversity of animal forms is proof of the range of variation that has evolved, it raises a question. Do animal forms fill up the space of possible forms or, more generally, does the genome have the potential to create an unbounded range of variation with no gaps? Answers to this question are only beginning to emerge, but they suggest that there are at least three factors that constrain the range of potential forms, creating gaps that have never been, and may never be, filled. Specifically, it may be highly improbable that animals evolve a particular form or class of forms — impossible morphologies — because of phylogenetic inertia (the tendency for a trait to remain stable in a variety of species because it was previously successful in a common ancestor), lack of relevant environmental pressures that result in selection among the biologically given options, and physical design constraints.

To illustrate these issues, consider two examples, one from molluscs and another from birds. Ammonoids are extinct cephalopod molluscs with a shell that spirals out from the centre before opening up. Studying the structure of their shells reveals two relevant dimensions that, when quantified, account for the observed variation (Fig. 1): the rate at which the spiral opens out, and the distance between the centre of this spiral and the opening[10,11]. If spiral rate is plotted against distance to the opening for the theoretically possible space of ammonoid species, as well as the actual space, there is a density of forms in a few areas and then gaps. The occupied spaces in this 'map' show the forms that evolved, whereas the vacant spaces suggest either possible morphologies that have not yet evolved or impossible morphologies.

The second example comes from the diversity of finch species living on the Galapagos Islands. Observations from the time of Charles Darwin to the present[12] show that interspecific differences in beak morphology among these finches evolved as a result of selective pressures from differences in seed morphology and availability. Like ammonoid shells, however, not all theoretically possible variants evolved, and nor would they be expected to, given various physical constraints, including those that involve producing song and maintaining head position during flight. Molecular studies have now uncovered the machinery that facilitated this variation[13,14]. Specifically, two genetic mechanisms control the length and height of the beak during ontogeny, by guiding the expression of genes that influence bone growth. For large-beaked finches, one of these proteins — known as bone morphogenetic protein 4 (BMP4) — is associated with bone growth earlier in development and is present at higher concentrations than in smaller-beaked finches. When the gene encoding BMP4 is inserted into a chicken

embryo, the developmental outcome is a chick with a large, broad beak, instead of the small beak that is typical of this species. Importantly, this experimentally induced, unnatural anatomical variant develops seamlessly, appearing much like the large-beaked Galapagos finch. This observation suggests that developmental programs are set up as 'engines of variation', providing a range of potential options for targeted selection. It also shows why cataloguing variation in living animals is insufficient for understanding both the range of variation and its potential constraints; experimental studies such as those with chickens are necessary to uncover the limits of variation.

The challenge in this research is to understand what causes gaps in form, harnessing the tools and theories of molecular biology, mathematics, physics, ecology, anatomy and behaviour. Why, specifically, have certain species never taken over a particular swathe of morphological turf? What is it about this space that leaves it vacant? Some of the causes will be intrinsic to the organism (such as the absence of genetic material or developmental programs for building square ammonoids), and some will be extrinsic (such as a lack of relevant ecological pressure that would favour hooked beaks).

Two central ideas emerge from studies of animal form. First, the observed differences, although adaptive, arise from a core set of cellular mechanisms that generate a massive range of possible variation[15]. Second, as generative as these mechanisms are, there are constraints, which arise from developmental factors, physics, history and ecology. As a result, to understand what a genome is capable of building, it is important to look not only to the fossil record or the extant animals but also to experimental manipulations that attempt to push nature beyond its visible bounds. Doing so will enable the causes of gaps in the range of potential animal forms to be better understood. Strikingly, this line of reasoning has direct parallels in the generative tradition of linguistics[16–19].

Gaps in linguistic form

Natural languages, both extant and extinct, are remarkably variable, seemingly boundless in terms of their variation in sound structures, lexicons and organizational principles. Inspired by early work in cellular biology[20,21], several linguists working in the generative tradition initiated by Noam Chomsky[17,22] started challenging the idea that there is unbounded variation in linguistic form. In particular, the earliest challenges suggested that the observed variation was highly constrained and mediated by a set of universal computations that enabled every developing human infant, but no other animal, to acquire a range of possible languages[23,24]. More specifically, the human brain has a uniquely evolved language capacity that links grammatical rules (syntax) with systems of meaning (semantics) and externalized expression in sound or sign (phonology) to provide a family of developmental options for building different languages[23–27]. Crucially, this perspective raised the idea of impossible languages: that is, linguistic structures that would either never be contemplated or, if contemplated and expressed, could not be learned[28]. For example, no language has a rule that mandates placing a particular word (for example, 'no') in a fixed position (for example, the fourth position) in a linear sequence. And no language stipulates that the words in a statement of fact (for example, 'The dog bites the man') are converted into a question by simply reversing the order of the words ('Man the bites dog the?'). Although there are controversies concerning the limits of linguistic variation, and the details of its universal structure, here I reveal three points of contact between work in the generative tradition of linguistics and evolutionary developmental biology research on animal forms.

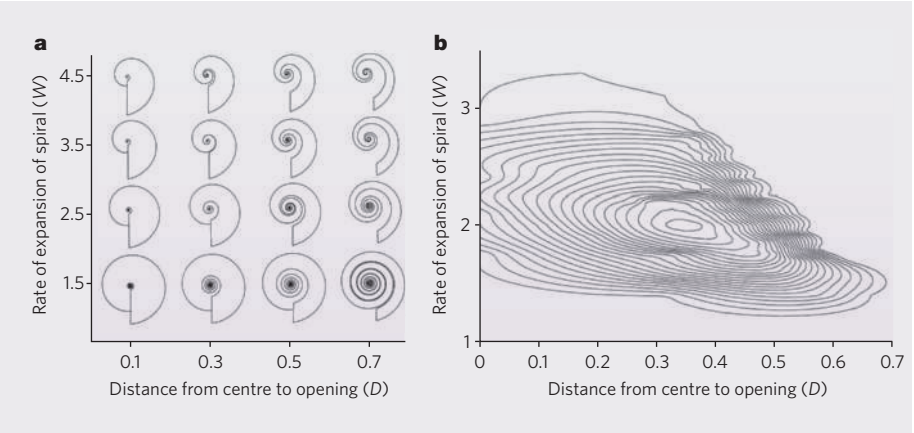


Figure 1 | The morphospace of ammonoid shells. The theoretical space of ammonoid shell forms (**a**) and the observed space (**b**) is shown. The rate of expansion of the spiral out from the centre (*W*) is plotted against the distance from the centre of the shell to the opening (*D*). Comparing the theoretical 'map' of possible forms (**a**) with the map of observed forms (**b**), it is clear that the upper right corner of the map of observed forms is empty, revealing that ammonoid shells of such forms have never evolved. (Figure reproduced, with permission, from ref. 10.)

A first point of contact is the fact that children are born with the capacity to acquire a wide range of possible languages, as opposed to specific languages such as English, Korean or French. This implies that a child is equipped with an abstract acquisition device, allowing the 'growth' of many different languages. Furthermore, as the child's acquisition device generates a space of possible languages, something internal or external to the device creates a space of impossible languages — forms that are never entertained by the child because they are poorly designed for acquisition and externalization in linguistic communication. In cases in which languages, such as Esperanto, are invented, they are acquired in a different way (for example, by relying on rote learning and teaching as opposed to spontaneous acquisition), and they prove unstable over the long term, dying out like the extinction of a species.

The beauty of thinking about the child's linguistic endowment as a system for building a space of languages is that it maps onto work in functional morphology described in the previous section[10]. Thus, in the same way that biologists speak of morphospaces — *n*-dimensional volumes that define the range of existing and potential morphological variation — linguists can speak of 'linguaspaces'. These are *n*-dimensional environments that constrain the set of possible languages and therefore, by definition, establish the set of impossible languages. What is necessary, therefore, is to establish the set of parameters that allow the range of variation and place constraints on its overall form. An elegant example, with direct parallels to language, comes from analyses of the design space of skeletal morphology. On the basis of a functional library of only seven parameters, each with a few options, it is possible to account for approximately 80% of the skeletal variation observed since the Middle Cambrian[29] (Fig. 2). When environmental conditions favour selection of a particular option for one parameter, this imposes constraints on the selection of options for other parameters.

Although it is not yet possible to define the dimensions that constrain the range of possible languages (and thus establish the features of impossible languages), adopting this perspective establishes an important methodological point. Akin to work in theoretical morphology, in order to understand the linguaspace, it is important to go beyond the extinct and extant languages — languages that have been generated — to languages that could be generated, acquired and maintained over time. Given human creativity, a group of people could create a novel language that violates many of the universally shared features of language (for example, massive embedding, no constraints on word order and unpredictable relationship between syntactic categories). Experimental study of this point might entail creating artificial languages that eliminate, or modify in significant ways, the set of linguistic universals that have been catalogued by linguists[30]. Given

such modifications, would individuals acquire this system as human children acquire known languages? And after the language has been acquired, would it remain stable, lasting across generations of users? And would such systems be represented in the same neural circuits that mediate natural languages? If the ideas laid out thus far are broadly correct, the answer to all three of these questions will be no.

A second point of contact concerns the nature of the input and the timing of growth and development. When a child 'grows' its native language, the appearance of certain structures is constrained by the appearance of others, as well as by the timing and magnitude of the input (that is, how much linguistic experience, at which time points, and from what sources). Some of these constraints are specific to language, and some result from the interaction between language-specific operations and processes that are more domain-general, including memory and learning. For example, recursive computations such as 'merge' (which, loosely defined, is an iterative operation that takes two elements and combines them into a set to create new expressions) are unlimited with respect to the number of iterated computations, but they are constrained by language-external processes of memory and comprehension, as well as by properties of the motor system that enable externalization, forced through a process of linearization (that is, each sound or sign has to emerge in communication in a linear sequence). By analogy, much of the work in evolutionary developmental biology suggests that the growth and development of different animal forms arises as a result of core molecular operations for generating variation (for example, rearrangement, repetition, magnification and division), with each of these processes further modified and constrained by the timing and magnitude of experience.

A third point of contact concerns how the internal language system ultimately forms an acquired and externalizable language. If, as discussed earlier, the acquisition device constrains the range of possible languages by providing a set of options, then the role of environmental input is to favour, and thus select, certain options over others. This selective perspective, although uncommon in the mind sciences, aligns more closely with other work in biology, including studies of the immune system[31], the development of animal forms, the wiring of neurons[32,33] and the acquisition of bird song[34]. For example, songbirds have evolved brains with a set of developmental options for creating variation in song-relevant acoustic forms. Depending on the environment, certain note types are selected and are then reproduced in particular orders to create population-specific dialects — and so it is for language acquisition by humans. When a child is exposed to a particular linguistic environment, the relevant linguistic input or experience fixes the available options to create an externalizable language that is comprehensible to those who will care for and compete with the child.

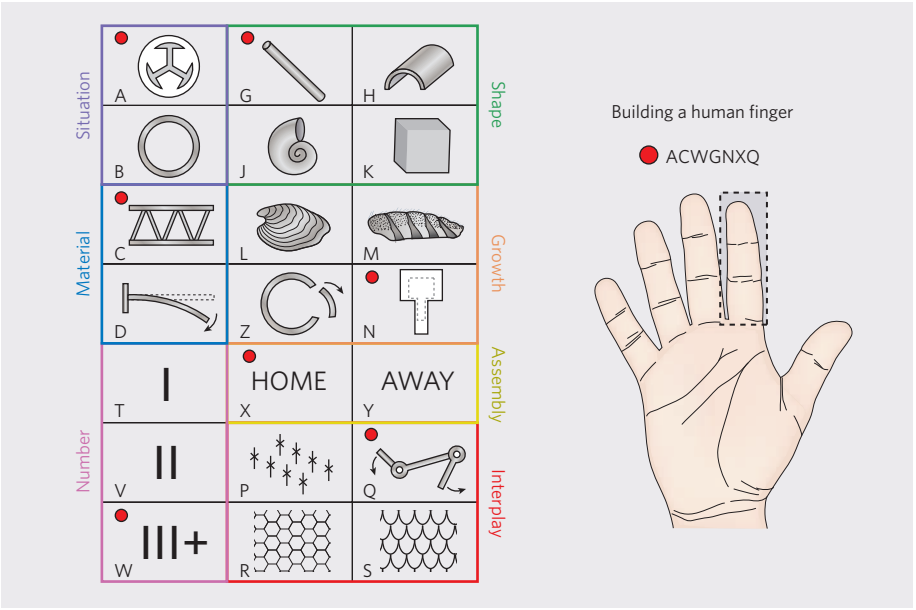


Figure 2 | The parametric space of skeletal variation. There are seven core properties or parameters (coloured borders), each with two to four possible states (each labelled with a capital letter), making a total of 21 variables. Situation indicates the location of the skeleton: internal (A) or external (B). Material indicates the composition of the skeletal elements: rigid (C) or flexible (D). Number refers to the number of skeletal elements: one (T), two (V) or three or more (W). Shape indicates the shape of these elements: rods (G), plates (H), cones (J) or solids (K). Growth indicates the way the elements are put together: by accretion (L), as branching serial units (M), by replacement or moulting (Z), or by remodelling (N). Assembly indicates where the elements are assembled: growth in place (X) or prefabrication (Y). And interplay indicates how the elements interact: no contact (P), jointed (Q), sutured or fused (R) or imbricate (S; that is, folded over or overlapping). A human finger is designed on the basis of one state (red circle) from each of the seven properties, specifically ACWGNXQ. (Figure modified, with permission, from ref. 29.)

Research in the generative tradition of linguistics suggests therefore that, like the variety of animal forms, the sense of unbounded variation in linguistic form is illusory, concealing a suite of universally held, biologically instantiated mechanisms for generating variation, allowing acquisition and constraining the space of possible languages. Although biologists have long sensed the close connection between the generative properties of language and generative biological systems, including the immune system, microbial diversity and proteonomics (the study of protein function and expression)[31,35,36], relatively few students of the mind sciences have acknowledged such connections with other domains of human knowledge[19,37].

Phylogenetic mind gaps

It has been argued that the history of life on Earth presents eight main evolutionary transitions, beginning with the replication of molecules to form populations of compartmentalized molecules and ending with the change from primate societies to human societies with language[38]. From this viewpoint, the birth of language caused a transition from non-human animals to humans. There are two problems with this view, however. First, it assumes, as Darwin did, that there is mental continuity among humans and other animals, while nonetheless acknowledging that language had a role in making humans different. Second, it treats language as a monolithic part of our psychology (as opposed to a capacity with a suite of distinctive computations), and it fails to recognize other, more distinctive, properties of brain function that facilitate the seemingly limitless variation in cultural expression and modes of thought. By contrast, I propose that humans evolved unique neural capacities after divergence from the last common ancestor some 6 million to 7 million years ago, and these capacities created a fundamental, and unprecedented, gap in the evolution of animal minds[39,40]. It created what I call our humaniqueness. Specifically, humans alone evolved four distinctive computational capacities.

Generative computation Recursive and combinatorial operations provide the only known mechanisms for generating an almost limitless variety of meaningful expressions, whether mathematical, linguistic, musical or moral. Recursion is an iterative operation, in which a rule is called up repeatedly to create new expressions, be they embedded phrases within a sentence, new musical scores with repeating themes, or tools within tools (for example, a Swiss army knife). Each expression has a unique interpretation or function depending on the arrangement of the elements. By contrast, combinatorial operations allow discrete elements to be unified and ordered, thus creating new ideas, which could be expressed as novel words (Walkman from walk and man) or novel musical forms.

Mental symbols Humans readily, without instruction, convert sensory experiences and abstract thoughts into externalized symbols, either as words or images. This capacity cuts across domains of knowledge and sensory experience, enabling humans to express beliefs in sentences, to depict particular melodies with explicit notations, and to provide logos indicating when to turn off the highway for a hamburger or a coffee.

Promiscuous interfaces Humans have unique creative capacities and problem-solving abilities, which stem from the capacity to combine representations promiscuously from different domains of knowledge. For instance, humans can combine the concepts of number, belief, causality and harm in deciding that it is sometimes morally obligatory to harm one person to save the lives of many.

Abstract thought Some thoughts derive from direct sensory experiences: for example, thinking of red items such as cherries and blood requires experience with these, as opposed to non-red objects such as celery and bone. But many human thoughts are abstract, with no explicit or even necessary sensory connection. These include concepts such as infinity, grammatical categories such as nouns and verbs, and ethical judgements such as permissible and forbidden.

The proposal that our humaniqueness, and these four properties in particular, finds almost no parallels in any other animal suggests that there were fundamental changes in genomic and neurobiological function during a relatively recent period of human evolutionary history. These changes provided humans with an unprecedented capacity to generate variation in culturally meaningful forms, albeit constrained in important ways.

The cultured gene

Although anthropologists disagree about the timing of the human cultural revolution[41–43], many researchers point to fundamental changes starting some 800,000 years ago in the Early Palaeolithic, with a crescendo of change at around 45,000–40,000 years ago in the Late Palaeolithic. This period is associated with the generation of symbols (mathematical, artistic and ritualistic), controlled fire for use in cooking and other forms of environmental transformation, and tools with multiple components and functions (for example, tools used for expressing both aggression and music). Given that this interval of several thousand years is barely noticeable on an evolutionary timescale, and that such cultural expressions emerged rapidly, the parallel with the Cambrian is striking: that is, something similar to a genetic revolution must have occurred during this period, providing humans with an unprecedented set of capacities for generating novel cultural expressions in language, morality, music and technology. Specifically, at some point before or during the Paleolithic, the human brain was transformed from a system with a high degree of modularity with few interfaces to a system of modules with numerous promiscuous and combinatorially creative interfaces. This system provided a universal framework on which cultural options could be realized[19,44]. To enrich these ideas, I next discuss comparative evidence that supports the evolutionary uniqueness of these four essential properties of human brain function and reveals the discontinuity — the gap — between human and animal minds.

Recursive and combinatorial operations are ubiquitous (and therefore domain-independent) in human mental life, operating in language, music, morality, technology and mathematics. A simple example, in the case of language, is creating a list, which has the recursive rule AND X+, where X is the name of a person. Thus, one list could be Sally AND Bill AND Sam AND Jane, and so on. This example, which every child immediately understands, illustrates the almost limitless capacity of humans to create linguistic expressions (that is, the property of discrete infinity), as well as the fact that the child's starting state is not blank but prepared with a competence that readily and implicitly understands recursive operations. In the example above, there is simply no experience that informs the child about the iterative and limitless power of list building.

Iterating a rule such as AND X+ is a type of looping operation in which the same computation is returned to repeatedly until some other function terminates the operation. Although many vertebrates have evolved brains with reciprocal connections or loops between different cortical areas (for example, basal ganglia to the cortex and back), these loops are restricted to particular functions[45,46]. At the most general level, it is clear that the motor systems of all animals must involve recursive operations to allow organisms to take a discrete set of motor options and generate a vast range of functionally meaningful motor acts or sequences in novel environments. For example, whether an organism flies or runs, its legs must repeatedly lift and fall or its wings must repeatedly beat. However, because an organism's habitat and climate is constantly changing, the iterative or recursive rule of cycling through leg lifts or beating the wings must be flexible so that the animal's response can vary in response to environmental change[47].

That said, the recursive properties of the motor system seem to be locked into motor function in all animals but humans. For example, in striking contrast to the recursive operations in human language, with its unrestricted use of different content or classes of words, the looping circuitry that is necessary for song acquisition in songbirds only supports singing and, in some cases, mimicry of other biological and non-biological sounds. This circuitry is not, however, used when they acquire the calls that constitute their repertoire more generally, including the sounds used in social interactions, food discovery and alarm calls.

Another example of generative computation comes from the domain of artefacts, in particular the creation and diversity of human tools. Unlike many of our simplest tools, such as the pencil, animal tools consist of a single material, never include more than one functional component, are typically dispensed after their first use and are never used for functions other than the original one. The first two features reveal that, unlike human tools, the representation of animal tools is not combinatorial. A pencil can combine four materials (graphite, wood, metal and rubber) to create four functions (graphite for writing, wood for holding the graphite, metal for attaching the rubber to the wood, and rubber for erasing). Moreover, each material can be used for a variety of other functions: for example, rubber can be a component of chewing gum. As experiments reveal, if a young child is asked what she can do with a pencil other than write, she will immediately offer such functions as holding up her hair, puncturing a plastic cover and poking a friend[48] (Fig. 3). Only humans think of artefacts as being designed for a particular function but, as a result of promiscuous interfaces, entertain many other possible functions.

The generative mechanisms that underpin so much of human mental life acquire their expressive power because the recursive and combinatorial operations can functionally 'grab' the outputs of different modular systems or domains of knowledge. This capacity for promiscuously creating interfaces between domains is almost absent in animals. Thus, although both human and animal brains are characterized by modular functions and mechanisms, the modular outputs are typically restricted to a single functional problem in animals but are broadly accessible in humans[44,49–51]. Non-human animals therefore show a form of myopic intelligence, designed to solve one problem with exquisite efficiency. For example, although honeybees have a symbolic dance that indicates the distance, direction and quantity of food, this communication system is largely restricted to food despite the intricate social lives of bees[52]. Although meerkat adults teach their pups how to kill scorpion prey by providing them with age-appropriate opportunities for handling and dismembering, teaching does not occur in any other context[53]. Although plovers use a deceptive display to lure predators away from their nest of eggs, they do not deceive in any other situation[54]. And although chimpanzees use the direction of another's eyes to guide strategic competition, they are far less skilled at using another's eyes to guide cooperation[55]. By contrast, in humans, neither language, teaching, deception, or the use of seeing to infer knowing are restricted to a single context.

Generative mechanisms in human cognition are enhanced further by the fact that they operate over mental symbols, reducing memory load

over non-symbolic representations. Furthermore, because symbols can be attached to any concept — whether real or imagined, abstract or anchored in sensory experience — they allow great flexibility in our modes of thought and communication. Although 40 years of research has been invested in the capacity of animals to produce or comprehend externalized symbols, the relevant evidence that they do so is, at best, weak, including findings from studies of natural communication or artificial acquisition of human-created signs or symbols[1,56–58]. Thus, even in cases in which animals spontaneously produce signals that seem to convey information about an external object (such as a predator or food) or event (such as movement of a group or submission to a dominant member), the number of different signals with such semantic properties is exceedingly small. The same limitations apply to studies of captive animals that have been trained to acquire a symbolic system[59]. These repertoires pale in comparison to even those of a human toddler, who can use hundreds of words and will soon have a repertoire of thousands.

Most of the conceptual representations acquired by animals seem to be largely anchored by sensory experiences, lacking the abstractness that characterizes so much of human thought. One exception to this comparative claim comes from recent studies of mathematical knowledge in humans and animals, in particular the neurobiological and psychological mechanisms that support both non-linguistic and linguistic quantification[2,44,60,61]. A wide range of animals are endowed with the capacity to quantify the number of individuals in an array not only spontaneously but also across modalities (for example, visual and acoustic), and even in cases in which they are required to integrate modalities.

Even with the abstractness of number representation in animals, however, there are important ways in which human thought, through its promiscuous interfaces, transforms this representation in unique ways. Consider, as an example, the distinction made in many human languages between singular and plural. At its core, this is a system of set-based discrimination that distinguishes between one and many but not between many and many. Thus, the morphological extension -s is used in English to count objects whether there are 2, 100 or 1 million cats, trees or pencils but not if there is only 1 of these objects. Studies of rhesus monkeys suggest that they share, with humans, an important aspect of this set-based quantificational system[62]. Specifically, in a foraging task in which sets of food items were presented and then concealed, rhesus monkeys preferentially picked a location with many pieces of food over a location with one piece, but not many pieces over many.

This non-linguistic, set-based system is present in human development but is then transformed as a result of an interface with the abstract properties of human syntax. Thus, although in English there are 2, 100 and 1 million cats, there are also −2, 0 and 1.0 cats. Consequently, when a neural connection was made between the evolutionarily ancient, set-based system and the evolutionarily recent syntactic system, a conceptual transformation emerged that was simple but abstract: anything that is not precisely 1 acquires the -s extension. This case study highlights both the limits of animal thought and the transformative effects on conceptual representation that emerge when abstract concepts from different domains of knowledge interface (Fig. 4).

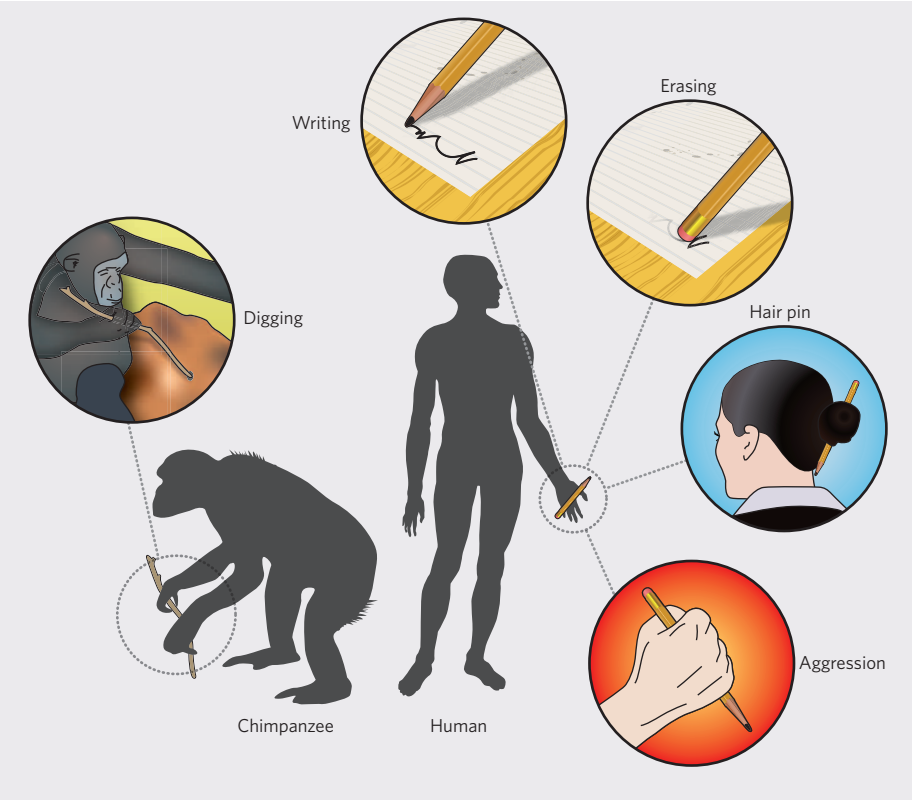


Figure 3 | Evolving the mind of a toolmaker. Whereas all non-human animals, including chimpanzees, use one object for one function, only humans have evolved the capacity to use one object for many functions, and to combine objects that each have a different functional role in order to solve novel problems.

Cultural gaps

Open a history book and explore the variation in visual arts, the changing attitudes about morally forbidden acts, the explosion of languages across the globe, and the endless gadgets that humans design to solve every problem from the most mundane (for example, a potato peeler) to the most profound (rockets to explore the Solar System). This shows a spectacular diversity with all the signatures of a system capable of change. No other animal mind is comparably endowed.

Controversy arises, however, over which cognitive capacities were crucial for creating the mental gap. And once the distinctive cognitive architecture of humans was in place, there is controversy over whether it provided humans with an unbounded potential to create cultural variation. I propose that much of the variation observed in human culture is highly constrained, with the space of possible cultures only sparsely populated, leaving several gaps that constitute impossible cultural forms. This perspective, with its parallels to work in theoretical morphology and extension of the general approach that has motivated work in generative linguistics, implies that some cultural forms will never be entertained or, if they are, will rapidly die out because they are unlearnable or learned with great difficulty. This view has interesting implications for both the study of culture and the biology (the genes, neural circuits and cognitive processes) that facilitates and constrains cultural acquisition and transmission.

The idea that there are cultural gaps raises the same kinds of questions as the idea that there are gaps in animal form. That is, it is necessary to understand what generates variation in cultural forms and why certain theoretically possible forms are never realized. On the basis of what can be observed, humans are born with a mental tool kit for creating, and especially for understanding, cultural variation in linguistic, musical, artefactual and moral expression. This tool kit consists of a suite of developmental programs that generate variation, the raw material for a selective process that crystallizes a particular form of expression. When cultural forms crystallize, perhaps as a result of a process of cultural inertia, gaps may emerge, because it is not possible for the individuals within the culture to imagine alternatives — a poverty of the imagination. In other words, people may get

stuck in a cultural rut because the human mind cannot move beyond the constraints imposed by history. The other main reasons for cultural gaps are the absence of relevant environmental triggers and the possibility that some ideas are never contemplated. These issues form the core of the argument for thinking about human cognition as massively generative but highly constrained. They also raise a serious methodological problem.

In many accounts of cross-cultural variation, there is a tendency to rely on ethnographies describing past and current cultures. But as in the case of animal forms, such descriptions showcase only what has been realized and not what could be realized in different circumstances or what might never be realized regardless of the circumstances. In particular, if humans are equipped with developmental programs that can generate a space of cultural expressions — a 'culturespace' — the observable cultures might occupy only a small fragment of the potential space. To uncover whether the currently empty space is within the range of theoretically possible cultures, experiments and computer simulations must be carried out to determine which of a range of possible cultural variants are evolvable, intelligible, acceptable, learnable and stable over time. Work like this is well under way in the study of language, in which researchers have implemented modelling, cognitive experiments on artificial grammars, and neuroimaging[28,63–65].

Some of these ideas have also begun to proliferate outside linguistics, in the less-studied domains of music[66–68] and morality[18], with intriguing experimental evidence and novel theoretical insights. Both music and language share core resources, such as the use of combinatorial operations and the representation of hierarchical structure — an analysis that has been supported by neuroimaging studies revealing that both of these engage common regions in the brain[69]. An important aim of this work is to map the range of possible musical forms, together with the mechanisms that generate and constrain such forms[67]. Significantly, this research refers not to what can be produced (as this is up to the whims of a composer and the quirkiness of people's preferences) but rather to what can be perceived. So are there musical forms that cannot be perceived and therefore represent impossible structures? The answer is, unambiguously, yes. For example, all music has a hierarchical structure in which there are groups of notes, phrases and sections. Therefore, a musical grammar that generates output in which grouping makes no difference represents a gap in the existing 'musicospace' and thus seems impossible. Furthermore, and paralleling the previously mentioned case in language of a grammar with a fixed position for a word class, imagine a musical system in which meaning is derived not from hierarchically related pitches but strictly from a fixed permutation of the set of 12 tones of the chromatic scale and from transformations on the entire set. This is a kind of structure that cannot be perceived except through conscious decoding[70].

An analogy has also been drawn between language and morality[18,19,71–74]. Evidence suggests that, like language, some of the computations underlying human moral judgements operate outside human awareness, are abstract, and show considerable similarity regardless of gender, age, religious belief or education[75,76]. Although, at present, there is no understanding of the possible and impossible moral systems — the 'moralspace' — the charting of this space is an inevitable outcome of the perspective taken here. For example, in several studies, including some cross-cultural work, individuals consistently judge that harming one individual as a means to some greater good (such as saving the lives of many) is worse than harming the one individual as a by-product or side effect. For example, if a toxic gas is about to be released through a vent into a room that contains ten people, it seems worse to push someone into the vent to block the toxic gas (killing one person but saving ten) than to press a button that shunts the ventilation into a second room where there happens to be just one person. From this work, it can be predicted that no moral system will ever operate on the reverse principle (that is, that side effects are worse than means). If this prediction is correct, a space of impossible moral systems, ones in which side effects would be judged worse than means, would be expected. And like the study of language and music, empirical inquiry into the range of possible moral systems is feasible given opportunities for modelling, carrying out cross-cultural experiments, and documenting normal and pathological brain function[77–79].

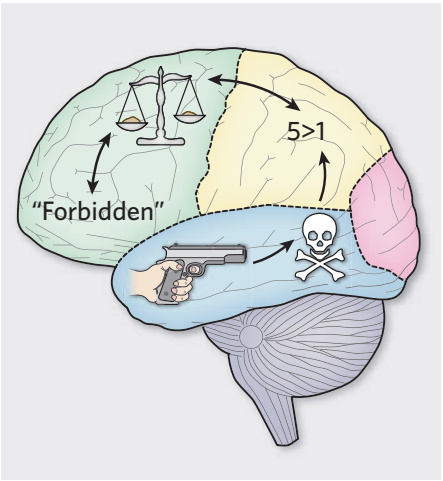


Figure 4 | Promiscuous interfaces between different domains of knowledge. A representation of an action (a finger pulling a trigger) interfaces with a representation of death as a potential consequence, which in turn interfaces with a system of numerical representation that evaluates whether the number of lives killed exceeds the number saved. This then interfaces with a moral evaluative system that judges the permissibility of the initial action, which then interfaces with the human linguistic system to deliver the judgement "forbidden".

Mind the gaps

The parallels I have drawn between the molecular biology of animal forms and the neurobiology of cultural forms will undoubtedly be incorrect in some details, but a science of the mind that focuses on the neural mechanisms that allow and constrain human generative systems is likely to make great progress. What is exciting is that some of the tools for exploring these questions are at hand or in development. For example, mice engineered to express the human version of the gene *FOXP2* (which has a role in speech production) learn motor skills (more specifically vocal behaviour) differently from mice with their own species-specific variant[80]. Similarly, neural chimaeras can be made, in which an area of one animal's brain can be inserted into another animal's brain, thereby changing the motor and behavioural capacities of the recipient[81]. Both genetic and neurobiological manipulations such as these allow researchers to go beyond cataloguing what animals can do and investigate what they might be able to do or not do. Cognitive scientists can now take advantage of these techniques to understand the actual, the possible and the impossible for different cultural expressions.

As exciting as these technological advances may be, they are relatively crude, especially given the questions being raised about cognitive evolution and the nature of human thought. We should not underestimate the great challenges ahead. For example, although language, music and mathematics recruit recursive and combinatorial operations to generate hierarchical structures, there has been no progress in identifying the circuits that are responsible for this ubiquitous capacity, including the genetic changes that made such domain-independent generative computations possible in humans but no other species. One might wonder, nonetheless, whether it will one day be possible to take the kinds of reciprocal or looping circuits observed in non-human animals, especially those that appear in all motor systems, and re-engineer them (either genetically or through neurodevelopmental manipulations) to interface with the conceptual resources that these species have evolved, creating a songbird that can sing not only with passion but also with the richly thematic and meaningful expressions that characterize a Wagnerian leitmotif.

Regardless of how far these techniques can be taken in the future, they open up unprecedented possibilities for understanding questions of evolution and cognitive capacity. Consequently, they show how the theories, technologies and findings of molecular biology, evolutionary developmental biology, neuroscience, cognitive psychology, linguistics and anthropology can be productively combined to understand one of the most profound problems of intellectual life: how humans evolved a uniquely generative brain that allows

massive cultural expression, and why there are gaps in the cultural record — spaces that no human brain will ever fill with linguistic, musical, moral or technological forms. ■

Marc D. Hauser is in the Departments of Psychology, Human Evolutionary Biology, and Organismic & Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
e-mail: mdh102559@gmail.com

1. Cheney, D. & Seyfarth, R. *Baboon Metaphysics* (Univ. Chicago Press, 2008).
2. Gallistel, C. R. *The Organization of Learning* (MIT Press, 1990).
3. Hauser, M. D. *Wild Minds* (Henry Holt, 2000).
4. Shettleworth, S. *Cognition, Evolution and Behavior* (Oxford Univ. Press, 1998).
5. Tomasello, M. & Call, J. *Primate Cognition* (Oxford Univ. Press, 1997).
6. Nisbett, R. E. *The Geography of Thought* (Simon & Schuster, 2004).
7. Shweder, R. *Why Do Men Barbecue?* (Harvard Univ. Press, 2003).
8. Tomasello, M. *Constructing a Language* (Harvard Univ. Press, 2003).
9. Nettl, B. *The Study of Ethnomusicology* (Indiana Univ. Press, 2005).
10. McGhee, G. R. *Theoretical Morphology* (Columbia Univ. Press, 1999).
11. Raup, D. M. *J. Paleontol.* **41,** 43–65 (1967).
12. Grant, P. & Grant, R. *How and Why Species Multiply* (Princeton Univ. Press, 2007).
13. Abzhanov, A. *et al. Nature* **442,** 563–567 (2006).
14. Abzhanov, A., Protas, M., Grant, B. R., Grant, P. R. & Tabin, C. J. *Science* **305,** 1462–1465 (2004).
15. Gerhart, J. & Kirschner, M. *Proc. Natl Acad. Sci. USA* **104,** 8582–8589 (2007).
16. Boeckx, C. & Piatelli-Palmerini, M. *Linguist. Rev.* **22,** 447–466 (2005).
17. Chomsky, N. *Rules and Representations* (Columbia Univ. Press, 1980).
18. Hauser, M. D. *Moral Minds* (Ecco, 2006).
19. Jackendoff, R. *Language, Consciousness, Culture* (MIT Press, 2008).
20. Jacob, F. *The Logic of Life* (Vintage, 1976).
21. Monod, J. *Chance and Necessity* (Collins, 1974).
22. Chomsky, N. *Syntactic Structures* (Mouton, 1957).
23. Jackendoff, R. *Foundations of Language* (Oxford Univ. Press, 2000).
24. Pinker, S. *The Language Instinct* (Morrow, 1994).
25. Anderson, S. R. & Lightfoot, D. W. *The Language Organ* (Cambridge Univ. Press, 2002).
26. Chomsky, N. *Language and Problems of Knowledge* (MIT Press, 1988).
27. Chomsky, N. *On Nature and Language* (Cambridge Univ. Press, 2000).
28. Moro, A. *The Boundaries of Babel* (MIT Press, 2008).
29. Thomas, R. D. K., Shearman, R. M. & Stewart, G. W. *Science* **288,** 1239–1242 (2000).
30. Lahiri, A. & Plan, F. in *Universals of Language Today* (eds Scalise, S., Magni, E. & Bisetto, A.) 31–58 (Springer, 2008).
31. Jerne, N. K. *EMBO J.* **4,** 847–852 (1985).
32. Changeux, J.-P. *Neuronal Man* (Pantheon, 1985).
33. Edelman, G. *Neural Darwinism* (Basic Books, 1990).
34. Marler, P. & Nelson, D. A. *Semin. Neurosci.* **4,** 415–423 (1992).
35. Gimona, M. *Nature Rev. Mol. Cell Biol.* **7,** 68–73 (2006).
36. Cases, I. & de Lorenzo, V. *Environ. Microbiol.* **4,** 623–727 (2002).
37. Pinker, S. *How the Mind Works* (Norton, 1997).
38. Szathmary, E. & Maynard Smith, J. *Nature* **374,** 227–232 (1995).
39. Penn, D. C., Holyoak, K. J. & Povinelli, D. *Behav. Brain Sci.* **31,** 109–178 (2008).
40. Premack, D. *Proc. Natl Acad. Sci. USA* **104,** 13861–13867 (2007).
41. Klein, R. G. *The Dawn of Human Culture* (Wiley, 2002).
42. McBrearty, S. & Brooks, A. S. *J. Hum. Evol.* **39,** 453–563 (2000).
43. Bar Yosef, O. *J. Israel Prehist. Soc.* **35,** 467–481 (2005).
44. Dehaene, S. in *From Monkey Brain to Human Brain* (eds Dehaene, S., Duhamel, J.-R., Hauser, M. D. & Rizzolatti, G.) 133–157 (MIT Press, 2005).
45. Striedter, G. F. *Principles of Brain Evolution* (Sinauer, 2005).
46. Edelman, G. M. *Neuron* **10,** 115–125 (1993).
47. Wolpert, D., Doya, K. & Kawato, M. *Phil. Trans. R. Soc. Lond. B* **358,** 593–602 (2003).
48. Keleman, D. *Trends Cogn. Sci.* **3,** 461–468 (1999).
49. Cheney, D. L. & Seyfarth, R. M. *How Monkeys See the World* (Chicago Univ. Press, 1990).
50. Mithen, S. J. *The Prehistory of the Mind* (Thames & Hudson, 1996).
51. Rozin, P. *Am. Behav. Sci.* **43,** 970–986 (2000).
52. von Frisch, K. *The Dance Language and Orientation of Bees* (Belknap, 1967).
53. Thornton, A. & McAuliffe, K. *Science* **313,** 227–229 (2006).
54. Ristau, C. in *Cognitive Ethology* (ed. Ristau, C.) 91–126 (Erlbaum, 1991).
55. Hare, B. & Tomasello, M. *Anim. Behav.* **68,** 571–581 (2004).
56. Call, J. & Tomasello, M. *The Gestural Communication of Apes and Monkeys* (Erlbaum, 2004).
57. Hauser, M. D., Chomsky, N. & Fitch, W. T. *Science* **298,** 1569–1579 (2002).
58. Savage-Rumbaugh, E. S. *et al. Monogr. Soc. Res. Child Dev.* **58,** 1–221 (1993).
59. Deacon, T. W. *The Symbolic Species* (Norton, 1997).
60. Cantlon, J. F. & Brannon, E. M. *J. Exp. Psychol. Anim. Behav. Process.* **33,** 32–41 (2007).
61. Nieder, A., Freedman, D. J. & Miller, E. K. *Science* **297,** 1708–1711 (2002).
62. Barner, D., Wood, J., Hauser, M. & Carey, S. *Cognition* **107,** 603–622 (2008).
63. Nowak, M. A., Komarova, N. L. & Niyogi, P. *Nature* **417,** 611–617 (2002).
64. Kirby, S., Cornish, H. & Smith, K. *Proc. Natl Acad. Sci. USA* **105,** 10681–10686 (2008).
65. Thompson, S. P. & Newport, E. L. *Lang. Learn. Dev.* **3,** 1–42 (2007).
66. Jackendoff, R. & Lerdahl, F. *Cognition* **100,** 33–72 (2006).
67. Lerdahl, F. *Tonal Pitch Space* (Oxford Univ. Press, 2001).
68. Lerdahl, F. & Jackendoff, R. *A Generative Theory of Tonal Music* (MIT Press, 1983).
69. Patel, A. *Music, Language and the Brain* (Oxford Univ. Press, 2008).
70. Lerdahl, F. *Contemp. Music Rev.* **6,** 97–121 (1992).
71. Dwyer, S. in *Philosophy and Linguistics* (eds Murasugi, K. & Stainton, R.) 169–190 (Westview, 1999).
72. Harman, G. in *Proc. 20th World Congress Philosophy* Vol. **1**. (ed. Brinkmann, K.) 107–115 (Philosophy Documentation Center, 1999).
73. Mikhail, J. *Trends Cogn. Sci.* **11,** 143–152 (2007).
74. Rawls, J. *A Theory of Justice* (Harvard Univ. Press, 1971).
75. Cushman, F., Young, L. & Hauser, M. D. *Psychol. Sci.* **17,** 1082–1089 (2006).
76. Hauser, M. D., Cushman, F., Young, L., Jin, R. K.-X. & Mikhail, J. *Mind Lang.* **22,** 1–21 (2007).
77. de Quervain, D. J.-F. *et al. Science* **305,** 1254–1258 (2004).
78. Henrich, J. *et al. Science* **312,** 1767–1770 (2006).
79. Koenigs, M. *et al. Nature* **446,** 908–911 (2007).
80. Groszer, M. *et al. Curr. Biol.* **18,** 354–362 (2008).
81. Balaban, E. *Proc. Natl Acad. Sci. USA* **94,** 2001–2006 (1997).

Acknowledgements I dedicate this paper to Noam Chomsky and Peter Marler for their insights into the many problems discussed here, as well as their friendship and advice over several years. For support during the writing of this article, and for research funds, I thank Harvard University's Mind/Brain/Behavior Interfaculty Initiative, J. Epstein, S. Shuman and the US National Science Foundation (Human & Social Dynamics priority area). For comments and discussion on drafts, I thank A. Abzhanov, A. Bejan, C. Boeckx, N. Chomsky, D. Dennett, B. Huebner, R. Jackendoff, N. Kanwisher, M. Kinsbourne, M. Kirschner, J. Junge, F. Lerdahl, G. McGhee, D. Roy, G. Striedter and F. Tsao.

Synthesis at the molecular frontier

Paul A. Wender and Benjamin L. Miller

Driven by remarkable advances in the understanding of structure and reaction mechanisms, organic synthesis will be increasingly directed to producing bioinspired and newly designed molecules.

Molecular evolution on Earth over the past 3.8 billion years has produced an extraordinary library of chemical structures, unsurpassed in number, diversity and function. Each structure is a treasure-trove of information and inspiration, a molecular textbook encoded in the language of chemistry. Collectively, these molecules comprise the chemical genome of our planet, and represent a universe ripe for exploration. With modern analytical tools, each of these structural tomes can now be read, enabling an understanding of how structure relates to function. More significantly, we can now use organic synthesis not only to make copies of these molecules, but also to prepare bioinspired or designed compounds, some with functions unheard of in the natural world — compounds that will influence, if not shape, every facet of our existence.

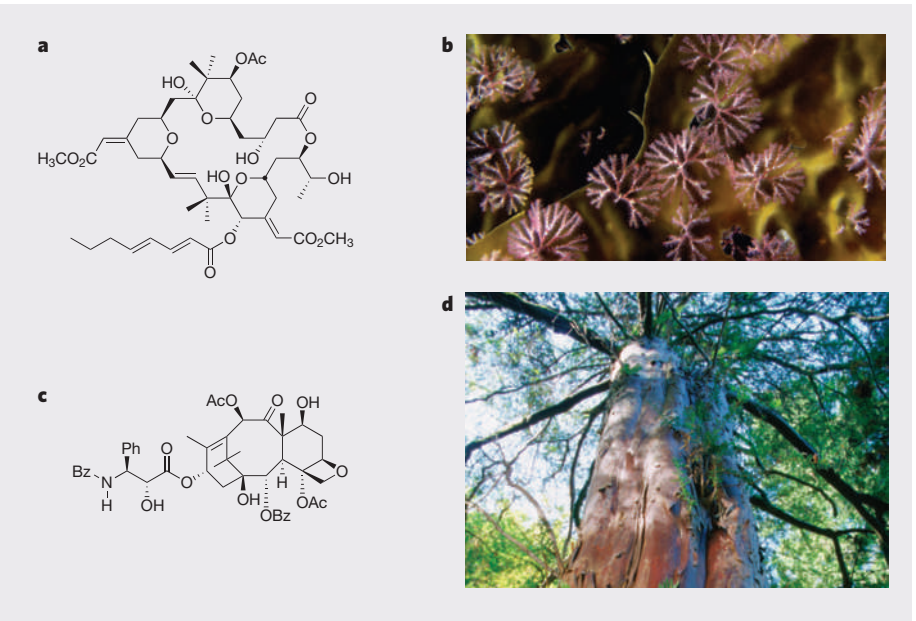
Our emerging molecular literacy is creating a revolution that will transform our world. The ability to design, create and control molecules has opened a vast frontier of research and an age of unprecedented opportunity. Scientists from every background are being drawn to this molecular frontier, creating a melting pot of disciplinary fusions and the resultant ability to address problems that transcend the boundaries of individual fields. From molecular biology to molecular computing, molecular medicine, molecular (nano) technology and even molecular gastronomy, science is becoming increasingly integrated and 'molecularized'.

Chemists have been laying the foundations for this molecular revolution for the past two centuries. Before that, nature's archive was the sole or primary source of chemicals used by humans. This has now changed. Through extraordinarily innovative advances in tools, theories and methods, synthesis has provided a reliable supply of many natural compounds as well as others created by design. Indeed the question of whether a molecule from nature could be made is increasingly giving way to whether it could be made in a way that impacts on supply and science. Of increasing importance now is the question of what molecules to make. Naturally occurring molecules are produced in their ecosystems for uses other than what we seek or need. Their activities in humans are thus serendipitous and unoptimized but provide a rich source of information and inspiration. We are now on the cusp of a period in which we can use this inspiration to design molecules with superior or new functions and make them in increasingly efficient, practical and environmentally friendly ways[1–11].

Interpreting the chemical genome

In 1861, the best that Alexander Butlerow could do to characterize the reaction product of aqueous formaldehyde and calcium hydroxide was to describe it as a golden liquid that tastes of liquorice. Today, molecular spectroscopy and microscopy allow us to detect the structural features of molecules, and even to observe their dynamic behaviour. To those schooled in the static representations of molecules in books, the movie of a single molecule moving in the barrel of a nanotube[12] inspires awe, along with a plethora of previously unimaginable ideas that span disciplinary boundaries.

Nuclear magnetic resonance (NMR) spectroscopy and single-crystal X-ray analysis have proven to be of particular importance to organic chemists historically, and are keys to the increasing ability to make known as well as designed functional molecules. In bygone times when spectroscopic methods for structural characterization were limited, natural products were targeted for synthesis — at least in part — because structural validation of the product could be made by comparing it to a natural sample. Making a non-natural, designed molecule 50 years ago would thus have carried the additional burden of establishing its structure. Analytical tools and techniques have now reduced or eliminated this problem for most molecules, thereby opening the door to molecular design.



I. SPENCE/VISUALS UNLIMITED/GETTY IMAGES; K. LEE

Figure 1 | Structural and functional inspirations from natural products. Bryostatin 1 (**a**) is putatively an antifeedant compound produced in the marine bryozoan *Bugula neritina* (**b**). This ornate natural product also has useful anticancer activity. Taxol (**c**) was isolated from the bark of the Pacific yew (**d**). It interferes with cytoskeletal dynamics to arrest the cell cycle and is used for the treatment of breast and ovarian cancers. Ac, acetyl; Bz, benzoyl; Ph, phenyl.

Box 1 | Concepts in synthesis

Total synthesis The production of a molecule from commercially or readily available chemicals through a series of chemical reactions.

Semi-synthesis The production of a molecule from a naturally derived starting material through a series of chemical reactions.

Convergent synthesis A total synthesis in which two (or more) molecular fragments are synthesized independently, then joined together.

Linear synthesis A total synthesis proceeding stepwise from commercially available starting materials, without the introduction of independently synthesized molecular fragments.

Ideal synthesis The production of a target molecule in a single synthetic operation from readily available starting materials, in 100% yield and without side-product formation. The synthesis should be simple, safe, economically acceptable and environmentally friendly.

Target-relevant complexity Bonds and stereochemical configurations that are retained in the final synthetic target.

Diversity-oriented synthesis The production of structurally diverse libraries of molecules. **P.A.W & B.L.M.**

Lessons from natural products

Nature's molecular archive encodes the evolutionary record of millions of species and of innumerable chemical entities: compounds that are part of complex biochemical systems involved in all aspects of life, reproduction and death. Some ants, for example, communicate through the use of hydrocarbons. The exchange of these chemical messages through antennal contact determines whether a recipient will harvest, forage or 'work at home' on a given day. In other organisms, chemical systems provide a sophisticated defence against predation. *Bugula neritina*, a marine bryozoan found along the Californian and other coasts, is proposed to be in a symbiotic partnership with a 'guest' bacterium in its gut. Together, they produce an agent known as bryostatin that is thought to serve as an antifeedant, protecting the bryozoan's larvae from predators (Fig. 1a)[13].

Remarkable as is bryostatin's natural role, the molecule also has other lessons for us, serving as a new lead for treating cancer[14] and, more recently, as an agent that facilitates learning and extends memory in animals[15]. Similarly, Taxol (paclitaxel; a compound found in the bark of the Pacific yew tree) arrests the cell cycle by interfering with cytoskeletal dynamics, thus teaching us much about cell proliferation and providing the collateral benefit of a highly effective anticancer drug (Fig. 1c)[16]. Indeed, many of the most important therapeutic agents introduced in the past 50 years are, or were inspired by, natural products[17]. More are surely to follow. Prostratin, for example, another plant-derived compound, is known to flush latent HIV from cells. Used in combination with antiretroviral therapy, this agent offers a promising strategy to eradicate HIV/AIDS[18,19].

In addition to therapeutic uses, many naturally occurring compounds have been the source of inspiration for the design of materials, diagnostics and imaging agents. Green fluorescent protein, for example, a molecule found in jellyfish, is now widely used for cellular imaging (see http://tinyurl.com/npe3r2) and the proteins and processes involved in photosynthesis have provided inspiration for solar collection devices[20], optical storage systems[21] and retinal prosthetics[22]. Clearly, we have much to learn from nature's molecular library.

Much like the exploration of our cosmic frontier, exploration of the molecular frontier has created vast new bodies of information. But a major difference between the two frontiers exists: not only can we observe the molecules of nature, we can also make them and, more significantly, even design and create new ones. Molecules that once were available only in trace amounts from natural sources can now be prepared and sometimes manufactured in quantity through total synthesis, semi-synthesis (Box 1), biosynthesis (using enzymes or even whole organisms) and combinations thereof. Synthesis in this way plays the metaphorical role of the printing press, creating copies of compounds that are otherwise scarce or found only in unreliable, inaccessible, or ecologically fragile natural sources. Complementing this capability, synthesis can also produce new molecules with improved or new functions, thereby expanding our frontier.

The ability to replicate nature's molecules has profound consequences. Not only can it be used to supplement supply, but the step-by-step synthesis of natural compounds also allows us to better understand how the parts of a molecule influence the properties and function of the whole — a form of reverse molecular engineering and thus inspiration on how it might be modified to create new medicines, materials, devices or probes. Throughout, the mantra of chemistry is repeated: structure, function, design and synthesis. Knowledge of structure leads to an understanding of function that informs design, leading to the synthesis of new structures with improved or totally new functions.

Bioinspired reaction science

Nature's cornucopia of diverse molecular structures has also served another profoundly important purpose: providing synthetic targets that inspire and often require the development of new reactions, methods, catalysts and strategies. Synthetic methods have evolved impressively over the past half century in response to this challenge. For example, a consideration of syntheses of the phorbol esters — remarkably active tumour promoters used in models of carcinogenesis — led to the design of a new metal-catalysed reaction for the synthesis of seven-membered rings (Fig. 2)[23]. A consideration of how nature introduces chemical groups into molecules has led to the design of dazzling 'C–H activation' strategies, in which ordinarily inert carbon–hydrogen bonds take part in reactions[24–26].

Whereas 50 years ago, literature reports of total syntheses of natural products were rare, such reports now appear daily, as a result of the many advances in reaction science. Numerous compounds, once considered to be beyond the reach of synthesis, are now produced synthetically in research or commercial quantities — even genomes are within reach of synthesis, in its broadest sense. This is both a testimonial to the impressive progress of the field, and a signal that it is time to take on the next level of challenges.

Not unlike the evolution of X-ray crystallography from studies on simple salts (such as sodium chloride) to studies on biomacromolecular complexes and even their dynamics, synthesis is poised to address a new range of problems, societal needs and scientific opportunities. Many — some would argue most — natural products can now be synthesized if suitable resources are provided. The challenge in synthesis is therefore increasingly not whether a molecule can be made, but whether it can be made in a practical fashion, in sufficient quantities for the needs of research and/or society, and in a way that is environmentally friendly if not 'ideal' (Box 1). Such syntheses require reactions that are more selective (minimizing the number of side-products) and efficient (providing good yields) than many of those currently available. But there are many more considerations.

Although selectivity and efficiency are important, their beneficial effects can be

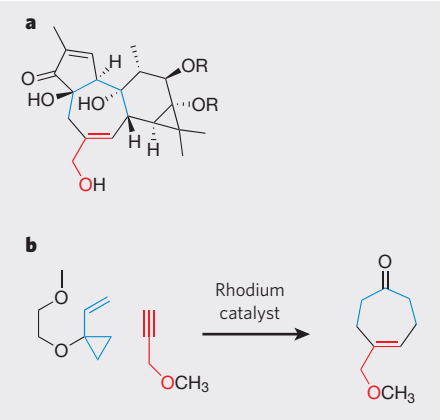


Figure 2 | The development of new reactions for step-economical total synthesis. Examination of the structures of phorbol esters (**a**) — remarkably active tumour promoters used in studies of carcinogenesis — inspired the development of a new rhodium-catalysed reaction for the synthesis of seven-membered rings (**b**).

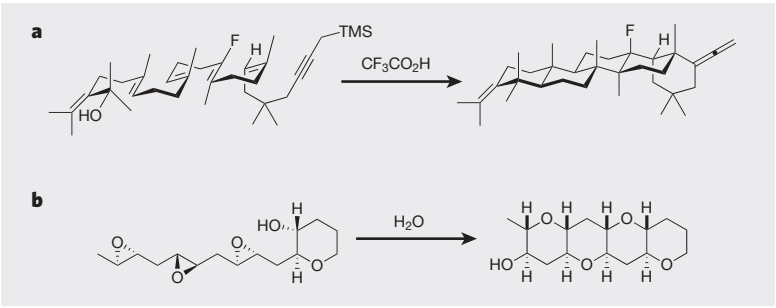


Figure 3 | The use of cascade reactions to rapidly generate complexity and value. **a**, A biomimetic complexity-building process involving carbocation intermediates (not shown) provides a powerful strategy for terpene and steroid synthesis. **b**, A related process that can be used to access complex polyethers via an epoxide-opening cascade initiated by water.

significantly compromised, or even offset, in syntheses that are too long to be commercially viable. For example, a synthesis of 70 steps, even if perfectly efficient (100% overall yield) and selective (no side-products), would be too expensive for most applications. This is because each step requires its own starting materials, reagents, solvents and energy costs, and produces a waste stream that requires disposal. More significantly, the biggest costs are in manpower: each step must be developed and executed by individuals skilled in synthesis. In contrast, shorter syntheses generally offer better economies, lower environmental impact, easier access to valuable synthetic targets and conservation of our most precious resource — those few skilled in the science and practice of synthesis. To appreciate the importance of 'step-economical' synthesis one need only reflect on how the one-step synthesis of carbon fullerenes (some of which are familiarly known as 'buckyballs') and nanotubes has enabled research around the globe in many different fields and has contributed to the emergence of nanotechnology.

The idea of 'step economy' is not new (although the term itself was coined recently)[27], but it is a singularly important issue in considering the current state and future of synthesis. A classic — and still spectacular — example of step economy is provided by a pair of syntheses of the alkaloid tropinone, a key intermediate in the production of atropine (a drug with many therapeutic applications). Tropinone was first synthesized in 1901 by Richard Wilstätter, in what was regarded at the time as a *tour de force* of organic chemistry. Even so, the synthesis required about 20 steps and had an overall yield of 0.75%. The compound was subsequently synthesized in one step in much better yield (17%) by Robert Robinson[28], who devised a multi-reactant, multi-bond-forming process inspired by biosynthetic reactions. Such dramatic improvements in syntheses are uncommon, but they call attention to the unique impact of new reactions, and the strategies they enable, on step economy. New reactions change the very way chemists think about a synthesis, and their views of what is possible. They create transformative rather than incremental change. The olefin metathesis reaction, in which two carbon–carbon bonds are broken and two new ones formed to create new products, is a recent example (see http://tinyurl.com/mkxmo6).

Towards the ideal synthesis

The step count of a synthesis is a function of which reactions are used and the sequence in which they are performed. Syntheses based on reactions that provide only small increases in target-relevant complexity (TRC; Box 1) generally require more steps than those based on reactions providing greater increases in TRC. The literature is replete with examples of this, as several different research groups succeed in synthesizing a given natural product, but with greatly differing step counts[29].

As most reactions involve the formation of only one or two bonds, one might conclude that syntheses of very complex compounds would invariably require many reactions and therefore many steps. There are, however, two ways around this apparent problem. The first entails performing a series of reactions in one vessel so that, as the products of each reaction form, they immediately go on to take part in the next reaction of the sequence, ultimately ending up in a final isolatable product. Because the products of each individual reaction are not isolated, the sequence constitutes a single synthetic operation in which many more than two bonds are formed and much time is saved, as only a single set-up, reaction and work-up is required. The second approach to step economy is to invent new reactions that generate greater increases in TRC per step. Both strategies shorten the journey from simple commercial materials to complex targets, with the ultimate goal being to arrive at an ideal synthesis (Box 1). Both save time, while minimizing cost and especially solvent use — solvents are usually removed and discarded at each step of a synthesis, and so are the greatest source of waste in most syntheses.

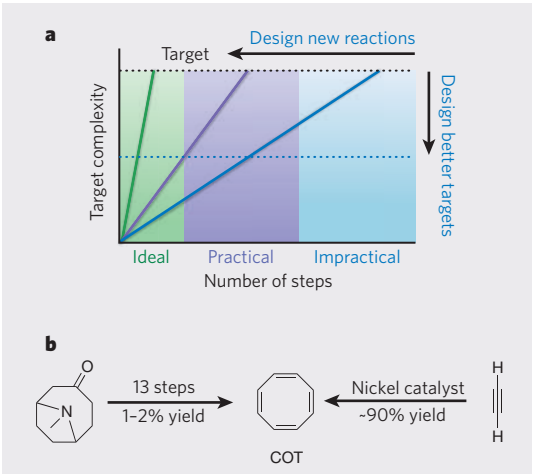


Figure 4 | Developing ideal syntheses. **a**, In an ideal synthesis (green line), a single step converts simple reactants into a structurally complex product. The syntheses of complex natural products often require too many steps (blue line), rendering them impractical for making large quantities of product. Reactions that generate a large increase in structural complexity per step allow shorter, more practical syntheses to be developed (purple line). For natural products that have valuable functional properties, another way to achieve step economy is to target structurally simpler compounds (blue dotted line) that have the same (or better) biological properties than the natural product. **b**, The first published synthesis of cyclooctatetraene (COT) required 13 steps, and the overall yield was very low. The discovery of a nickel-catalysed reaction allowed COT to be made in 90% yield in one step.

Such single-operation, multiple-step processes — described variously as serial, cascade, chain, tandem, regenerative or domino reactions — are an increasingly important focus of many research groups[30], and are certainly having an impact on the ability of synthesis to produce complex targets in a step-economical fashion. This strategy draws inspiration, if not validation, from biosynthetic pathways that often proceed through such serial processes. For example, the biosynthetic route to terpenes and steroids has been mimicked with great success through the efforts of Bill Johnson and others (Fig. 3a)[31]. Johnson's synthesis features a biomimetic complexity-building step in which a carbocation intermediate is generated; this is used to make a bond in a way that generates a second carbocation, and the cycle is repeated until the process is terminated. This is a wonderful example of intermediate recycling (regeneration). A related process (Fig. 3b)[32] that builds on a long-standing biosynthetic hypothesis allows access to complex polyether molecules (found in many marine natural products). Similar cascades have been reported that involve other reactive intermediates, including anions, carbenes[3] and radicals[33]. Even cross-over processes, in which one reactive intermediate

produces a second, different kind of intermediate[34], are emerging as innovative features of step-economical syntheses.

The chain reactions used to make polymers and oligomers from monomers are also obvious examples of successful cascade processes that involve the regeneration of reactive intermediates. An initiator I reacts with a monomer A to produce an intermediate I–A, which reacts with another monomer to produce I–A–A and on to the final I–$[A–A]_n$ product. Although promising[35], this approach is not yet applicable to a broad range of targets, because of the still unsolved challenges of sequencing the reaction of different monomers (A, B, C) to make, say, I–A–B–C rather than I–A–C–B. Nature choreographs such feats through compartmentalization, proximity and templating processes. Mimicking these processes in the lab represents a priority research opportunity for chemists.

So what of the second approach for achieving greater step economy: the design or discovery of new reactions that greatly increase TRC? As we have seen, such reactions offer tremendous opportunities to move syntheses towards the ideal (Fig. 4a). A historical example is the synthesis of cyclooctatetraene, a relatively simple-looking compound of significant theoretical interest, which originally required more than 10 steps to produce in a 1–2 % overall yield (Fig. 4b). Although the efficiency and selectivity of this synthesis could no doubt be optimized further, the introduction of a new reaction — a nickel-catalysed process in which four acetylene (H–C≡C–H) molecules react to form four bonds in one step — allowed cyclooctatetraene to be made in one step, in more than 90% yield[36]. One of the more striking examples yet of a complexity-generating reaction is the arene–alkene metaphotocycloaddition, in which three new bonds and up to six chiral carbon centres are formed, an exceptional increase in complexity. This reaction has greatly shortened the syntheses of numerous natural products, underscoring the profound impact that new complexity-generating reactions can have on step economy[37].

Despite impressive progress in the field of synthesis, we are currently awash with complex, potentially useful natural products that cannot be made in the quantities required for scientific research or therapeutic use. In some cases, enzymes, cells, or even whole organisms can be used to produce such compounds, simple analogues of them, or precursor materials easily converted to the compound of interest[38]. This approach benefits from nature's refined pathways, but the specific nature of biosynthetic machinery means that there are limits to the range of structures it can prepare. For fields such as drug discovery, this limitation suggests that, in addition to natural products and natural-product-like compounds available through biosynthesis, one must embrace a broader range of structural (and synthetic) possibilities. New molecular scaffolds are needed if we are to truly exploit the potential of organic synthesis.

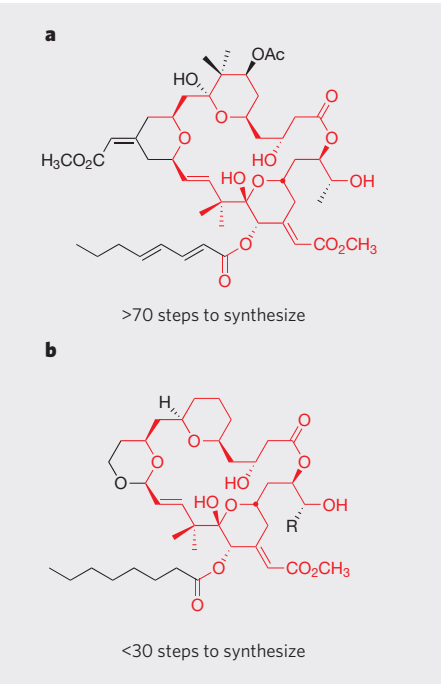


Figure 5 | Function-oriented synthesis as applied to bryostatin. Currently available syntheses of bryostatin 1 (**a**) involve more than 70 steps, meaning that they will have no impact on supply and are as yet impractical for industrial-scale production of the compound. An analysis of the complex structural features of bryostatin suggested which parts of the molecule are essential for the compound's biological activity and which parts are not. This allowed a structurally simpler analogue of bryostatin 1 to be designed (**b**), which was made in fewer than 30 steps. In cell and animal assays, this compound (and related designed analogues) performed as well as bryostatin in binding to cellular protein targets and in arresting the growth of cancer cells. Bonds and atom labels shown in red have been used to highlight parts of the two molecules that are identical.

Some parts are greater than the whole

It is now recognized that, in many cases, only certain parts of naturally occurring molecules are needed for functional activity[27]. This expands the focus and challenge of synthesis from making a known, complex structure, to designing a simplified but functionally superior target that might be more readily prepared. Bryostatin, for example, has been tested in clinical trials as an anticancer agent and is a promising candidate for drug-discovery programmes targeting cognitive dysfunction and Alzheimer's disease. But realization of the full clinical potential of bryostatin has been hampered by its limited supply: natural sources have very low levels of bryostatin, and total syntheses of the clinically relevant, potent bryostatins are at present too long to be a source of supply (exceeding 70 steps)[39].

Although practical access to bryostatin could eventually be addressed by optimizing and innovating around the existing synthetic approaches, it is important to remember that bryostatin, like most of nature's compounds, is neither produced nor optimized for human use. The design of structurally simpler but functionally superior analogues — an approach dubbed 'function-oriented synthesis'[27] — could be faster than optimizing syntheses of intricate natural products. Significantly, function-oriented synthesis also offers a way to tune, if not totally redesign, promising natural materials for optimal function.

In 1986, we analysed the structural features that contribute to bryostatin's unique biological activity, and used this analysis to design simplified bryostatin analogues that are easier to make than the natural product (Fig. 5). In cell and animal assays, many of these analogues performed as well as (and often better than) bryostatin in binding to cellular protein targets and in arresting the growth of cancer cells[14]; a second generation of compounds performed even better[40]. Crucially, these analogues can be prepared in a convergent synthesis (Box 1) that requires fewer than 30 steps, more than 45 steps fewer than required to make the clinically relevant natural product. This step-economical and tunable route to highly potent bryostatin analogues now drives research and preclinical studies on these compounds in drug-discovery programmes.

Similarly, the anticancer activity of the dimeric alkaloid stephacidin B can be mimicked by a monomeric alkaloid, avrainvillamide, which in turn can be mimicked by much simpler analogues[41]. Synthetic studies on the highly complex molecule halichondrin B led to the discovery that its antitumour activity was contained in one part of the structure[42]. A highly simplified analogue was then made that has even better activity than the natural product. The supply problem for halichondrin B has thus been solved by developing an efficient synthesis of a simpler analogue, which has entered human trials. Function-oriented synthesis can also be used to reprogram the properties and biological activities of compounds, as shown with amphotericin[43], Taxol[44], DNA binders and many others.

Function-oriented design and synthesis offers several advantages when compared with target-oriented synthesis. First, the natural product is not necessarily seen as the synthetic target, but as a source of structural and functional information that informs design. Traditionally, natural products were often selected for synthesis because their structures were previously established when isolated, frequently through lengthy research. Chemists would devise a route to make the proposed structure of a natural product, and if the compound thus obtained was identical to that isolated from a natural source, then the structure of the synthetic material and its precursors were validated. With some exceptions, modern analytical techniques have virtually eliminated the

need to establish structure by comparison with a structurally characterized natural product. This allows one to make designed molecules with confidence in their structures and the precursor structures leading to their syntheses.

So in addition to targets served up by nature, chemists can now design their own targets — molecules that have improved or new functions compared with their natural counterparts. This change of emphasis — from producing known structures to translating them into functionally superior but simplified structures — accounts in part for the recent explosion of non-natural targets and opportunities in synthesis. The rapidly growing interest in function provides a new range of exciting targets that have immediate relevance and application to many other areas of science.

Form follows function

Focusing on designed targets allows one to address a wider range of questions beyond those posed by the natural product, and often yields faster and superior results. Of course, natural products that incorporate unusual structural features are also justifiably recognized as inspiration for synthetic innovation. Function-oriented synthesis can also be used for this purpose, and often to better advantage, as one can design the target to fit both functional and synthetic goals. For example, designing a molecular muscle — a molecule that shortens or elongates electrochemically — poses a rich set of new synthetic challenges not unlike those encountered with natural products. The same is true of many other functionally desirable targets, such as smart molecules that detect (or prevent) disease, report on molecular events or interfere with biological processes. All of these could transform therapy and profoundly reduce the burgeoning cost of health care. Designed targets also often allow one to rapidly test hypotheses about the structural foundations of chemical reactivity, and to develop or improve synthetic methods, and the translation of a natural-product structure into a simplified, but functionally superior, designed compound is itself a great challenge to creativity.

There are many other opportunities for function-oriented synthesis, both in chemistry and in the growing list of molecularized sciences. As we have seen, synthesis is already enhanced by the use of biosynthetic machinery; synthetic biology represents an emergent field that takes the idea of biosynthesis and inverts it, by using synthesis to create new types of biology[45]. Diversity-oriented synthesis (Box 1) specifically targets non-natural compounds with natural-product-like complexity[46], and has identified small molecules with biological activities not yet displayed by known natural products[47]. Dynamic combinatorial chemistry is another synthesis-driven discovery concept that incorporates aspects of the ideal synthesis. It is a process in which a structure is selected and amplified from an equilibrating mixture of components on the basis of a desirable molecular property, for example binding to a biological target[48]. Recent efforts combining structural design (inspired by natural products) with dynamic combinatorial chemistry have led to the discovery of molecules that target RNA sequences critical to HIV replication[49], and to the most common form of adult-onset muscular dystrophy[50]. Molecular self-assembly and directed molecular evolution offer yet more opportunities to those with skills in synthesis and knowledge of function.

Conclusion

Synthesis is increasingly driven by an interest in molecules that not only have intriguing structures, but also have special functions. Added to this mix are exciting opportunities for designing molecular systems that go far beyond any natural function: from barcodes to electronic devices to self-assembled replicating systems.

Over the past two centuries, chemistry and synthesis have evolved from relatively pure disciplinary pursuits to positions of central importance in the physical and life sciences. More generally, they have provided the language and methodology that have unified, integrated and, indeed, molecularized the sciences, shaping our understanding of our molecular world and the direction, development and destiny of scientific research. From new reagents, reactions, strategies and processes to materials, medicinals and machines, chemistry and synthesis have shaped, and been shaped by, science and societal needs. Their continuing importance as central rather than as service sciences is partly a function of critical introspection (that is, can we do better?) but largely a function of sensing new opportunities in science of consequence to society. The human enterprise was once largely limited to what nature produced. We are now limited only by what we can imagine. ■

Paul A. Wender is in the Chemistry Department and, by courtesy, the Department of Chemical and Systems Biology, Stanford University, 333 Campus Drive, Stanford, California 94305-5080, USA. Benjamin L. Miller is in the School of Medicine and Dentistry, University of Rochester, 601 Elmwood Avenue, Box 697, Rochester, New York 14642, USA.
e-mails: wenderp@stanford.edu; Benjamin_miller@urmc.rochester.edu

1 Seebach, D. *Angew. Chem. Int. Edn* **29,** 1320–1367 (1990).
2. Cornforth, J. W. *Aust. J. Chem.* **46,** 157–170 (1993).
3. Wender, P. A. & Miller, B. L. in *Connectivity Analysis and Multibond-forming Processes in Organic Synthesis: Theory and Applications* (ed. Hudlicky, T.) 27–66 (JAI Press, 1993).
4. Tietze, L. F. *Domino Reactions in Organic Synthesis* (Wiley-VCH, 2006).
5. Nicolaou, K. C. *Tetrahedron* **59,** 6683–6738 (2003).
6. Wilson, R. M. & Danishefsky, S. J. *J. Org. Chem.* **71,** 8329–8351 (2006).
7. Horvath, I. & Anastas, P. T. *Chem. Rev.* **107,** 2169–2173 (2007).
8. Corey, E. J., Czakó, B. & Kürti, L. *Molecules and Medicine* (John Wiley, 2007).
9. Hudlicky, T. & Reed, J. W. *The Way of Synthesis: Evolution of Design and Methods for Natural Products* (Wiley-VCH, 2007).
10. Trost, B. M. *Science* **254,** 1471–1477 (1991).
11. Maimone, T. J. & Baran, P. S. *Nature Chem. Biol.* **3,** 396–407 (2007).
12. Koshino, M., Solin, N., Tanaka, T., Isobe, H. & Nakamura, E. *Nature Nanotechnol.* **3,** 595–597 (2008).
13. Sudek, S. *et al. J. Nat. Prod.* **70,** 67–74 (2007).
14. Wender, P. A. *et al.* in *Drug Discovery Research* (ed. Huang, Z.) 127–162 (John Wiley, 2007).
15. Sun, M.-K. & Alkon, D. L. *Eur. J. Pharmacol.* **584,** 328–337 (2008).
16. Kingston, D. G. & Newman, D. J. *Curr. Opin. Drug Discov. Dev.* **10,** 130–144 (2007).
17. Newman, D. J. & Cragg, G. M. *J. Nat. Prod.* **70,** 461–477 (2007).
18. Wender, P. A., Kee, J. M. & Warrington, J. M. *Science* **320,** 649–652 (2008).
19. Johnson, H. E., Banack, S. A. & Cox, P. A. *J. Nat. Prod.* **71,** 2041–2044 (2008).
20. LaVan, D. A. & Cha, J. N. *Proc. Natl Acad. Sci. USA* **103,** 5251–5255 (2006).
21. Hampp, N. *Chem. Rev.* **100,** 1755–1776 (2000).
22. Greenbaum, E. *et al. Tech. Dig. Int. Electron. Device Meet.* 496–498 (2002).
23. Wender, P. A. *et al. Pure Appl. Chem.* **74,** 25–31 (2002).
24. Hinman, A. & Du Bois, J. *J. Am. Chem. Soc.* **125,** 11510–11511 (2003).
25. Chen, M. S. & White, M. C. *Science* **318,** 783–787 (2007).
26. Chen, H., Schlecht, S., Semple, T. C. & Hartwig, J. F. *Science* **287,** 1995–1997 (2000).
27. Wender, P. A., Verma, V. A., Paxton, T. J. & Pillow, T. H. *Acc. Chem. Res.* **41,** 40–49 (2008).
28. Robinson, R. *J. Chem. Soc.* 762–768 (1917).
29. Chanon, M., Barone, R., Baralotto, C., Julliard, M. & Hendrikson, J. B. *Synthesis* **1998,** 1559–1583 (1998).
30. Wender, P. A. (ed.) *Chem. Rev.* **96,** 1–600 (1996).
31. Yoder, R. A. & Johnston, J. N. *Chem. Rev.* **105,** 4730–4756 (2005).
32. Vilotijevic, I. & Jamison, T. F. *Science* **317,** 1189–1192 (2007).
33. Curran, D. P. & Kuo, S.-C. *Tetrahedron* **43,** 5653–5661 (1987).
34. Padwa, A. & Bur, S. K. *Tetrahedron* **63,** 5341–5378 (2007).
35. Nicewicz, D. A., Satterfield, A. D., Schmitt, D. C. & Johnson, J. S. *J. Am. Chem. Soc.* **130,** 17281–17283 (2008).
36. Reppe, W. *et al. Liebigs Ann. Chem.* **560,** 1–92 (1948).
37. Chappell, D. & Russell, A. T. *Org. Biomol. Chem.* **4,** 4409–4430 (2006).
38. Das, A. & Khosla, C. *Acc. Chem. Res.* **42,** 631–639 (2009).
39. Hale, K. J., Hummersone, M. G., Manaviazar, S. & Frigerio, M. *Nat. Prod. Rep.* **19,** 413–453 (2002).
40. Wender, P. A., DeChristopher, B. A. & Schrier, A. J. E. *J. Am. Chem. Soc.* **130,** 6658–6659 (2008).
41. Wulff, J. E., Herzon, S. B., Siegrist, R. & Myers, A. G. *J. Am. Chem. Soc.* **129,** 4898–4899 (2007).
42. Towle, M. J. *et al. Cancer Res.* **61,** 1013–1021 (2001).
43. Szpilman, A. M., Manthorpe, J. M. & Carreira, E. M. *Angew. Chem. Int. Edn* **47,** 4339–4342 (2008).
44. Dubikovskaya, E. A. *et al. Proc. Natl Acad. Sci. USA* **105,** 12128–12133 (2008).
45. Krueger, A. T. & Kool, E. T. *Chem. Biol.* **16,** 242–248 (2009).
46. Schreiber, S. L. *Science* **287,** 1964–1969 (2000).
47. Thomas, G. L. *et al. Angew. Chem. Int. Edn* **47,** 2808–2812 (2008).
48. Corbett, P. T. *et al. Chem. Rev.* **106,** 3652–3711 (2006).
49. McNaughton, B. R., Gareiss, P. C. & Miller, B. L. *J. Am. Chem. Soc.* **129,** 11306–11307 (2007).
50. Gareiss, P. C. *et al. J. Am. Chem. Soc.* **130,** 16254–16261 (2008).

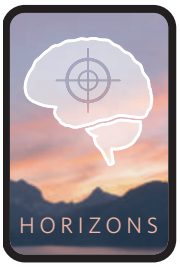
SIMON WATSON/TAXI/GETTY

Can biological indicators predict the future mental health of our children?

Biomarkers in psychiatry

Ilina Singh and Nikolas Rose

The use of biomarkers to predict human behaviour and psychiatric disorders raises social and ethical issues, which must be resolved by collaborative efforts.



Psychiatry has long been a second-class citizen in science and medicine. Despite much effort, the causes of many psychiatric disorders remain unclear, and it has been difficult even to categorize such disorders precisely. In the past decade, however, there has been a large shift towards incorporating biomarkers into psychiatry (Fig. 1), and there is hope that such biological indicators will improve psychiatric diagnoses by underpinning them with physiological evidence (Boxes 1 and 2). But biomarkers promise far more than a basis for better diagnoses. They could assist in predicting the course of an illness in an individual and in tailoring treatment. And they could be used to predict the development of not only psychiatric disorders but also certain behaviours, personality traits and mental or emotional capacity.

Scientific innovations that will ultimately improve psychiatric outcomes and general well-being are to be welcomed. But they must be scrutinized to assess their value to the general public. Despite the wealth of research into biomarkers and the considerable interest in their use in clinical and non-clinical situations, there has been little discussion of the social, ethical and legal problems posed by their use in psychiatry. Here we set out the key challenges in this area. We focus on interventions in children and adolescents, particularly those aimed at preventing behavioural problems. The identification of biomarkers in these age groups forms an important research agenda and the initial pathways through which this research is being translated from the laboratory to the clinic, as well as the classroom and other locations, can already be observed.

The promise of biomarkers

At present, psychiatric disorders are diagnosed on the basis of signs, symptoms and course of illness, according to the classifications in the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV). Finding a biological or physiological marker, rather than relying on behavioural symptoms and signs, might provide a more precise means of diagnosis, thereby aligning psychiatric classification with classification systems used in other areas of medicine[1]. Such methods might also go further and help to re-organize the DSM system of classification, offering a counter to the swelling catalogue of categories, each with its lengthening lists of behavioural symptoms and subclassifications that have no differentiated aetiology.

Moreover, biomarkers might be used to predict the potential for developing a particular disorder. This is of particular significance in child and adolescent psychiatry. Genetic screening and neuroimaging — the main techniques for identifying biomarkers — could be used to assess children before symptoms appear. And existing childhood disorders are now themselves being viewed as 'biomarkers' for the risk of developing more severe disorders.

In this sense, biomarkers promise to be the most powerful psychiatric tool since the discovery of antipsychotic drugs — a biological means of predicting not only the development of a disorder but also its course and outcome. Biomarkers could therefore inform the type, timing and course of interventions, and they could allow disorders to be subtyped based on physiological criteria, creating a more personalized approach to psychiatric treatments.

But the potential impact of psychiatric biomarkers extends beyond the clinic, to arenas responsible for the growth and development of children as productive citizens: the classroom and the courtroom. For example, in the United Kingdom, a new national agenda of ensuring 'mental capital and well-being' is grounded in new science that has transformed the understanding of child development and learning[2]. Enthusiasts for educational programmes based on neuroscience argue that neuroscientific evidence should be applied at an early age and with a broad remit, not only informing how children are taught and how classrooms are structured but also helping to identify developmental challenges — such as impulsiveness and learning difficulties — that are thought to be associated with later psychiatric, educational and social problems[3]. In the courtroom context, neuroscientific understanding of how the brain develops is being used to inform juvenile justice decisions. A landmark 2005 decision in the US Supreme Court that overturned the death penalty for juveniles (those under 18 years of age) is thought to have been strongly influenced by neuroscientific evidence about the capacities of the 'adolescent brain' to control impulsive and risk-taking behaviours[4]. In addition, research efforts are underway to identify brain-based biomarkers associated with juvenile delinquency so that neurodevelopmental risks can be built into models to predict youth antisocial behaviour[5]. These non-clinical applications of psychiatric biomarkers suggest the extent to which biomarkers could come to shape the lives of 'normal' individuals, especially children.

The challenges

The use of biomarkers in any of these contexts presents many challenges. A biomarker — for example, a certain pattern of brain activity — is not the cause of a disorder. So the current interest in biomarkers is a sign that psychiatry has undergone a methodological shift, away from searching for the causes of a condition towards estimating the probability that the condition is present or will develop. One challenge arising from this approach is that individual variables associated with an increased risk of developing a condition, for example information about a single biomarker, usually have small effect sizes: that is, when used alone, they are not robust predictors of the presence or absence of a condition. However, when bundled into an algorithm that incorporates other biomarker information, as well as social and environmental risk factors (such as prenatal exposure to alcohol, postnatal exposure to lead, family poverty and child abuse), biomarkers could be powerful predictors that a disorder will develop. However, biomarker information will always be a statement of the probability that a condition will develop: that is, a statement about the risk of developing that condition. Therefore, even risk algorithms that are powerful predictors will retain a degree of uncertainty. The hope is that this uncertainty will lessen over time, as highly predictive variables become easier to identify.

Many psychiatric researchers, however, have deep-seated doubts about whether the current methods for identifying biomarkers — such as genome-wide association studies (which search for genetic markers associated with disease risk) and imaging of the brain region of interest, which are based on studies of groups — can uncover biomarkers with strong predictive value for a specific individual. There are also doubts that biomarkers will have translational applications within the next decade in a manner that would allow biomarker information to guide clinical, educational and legal practices and policies substantively.

Outside specialist circles, these doubts are often minimized, and questionable biomarkers can begin to take on scientific and social importance. This process of overgeneralization and oversimplification has occurred on many occasions, for a long list of neuroscientific discoveries, including claims about the 'gene for aggression' in humans and about the implications of 'mirror neurons' in non-human primates (neurons that are active when an animal carries out a particular action or observes another animal carry out the same action), which have been equated with these animals having human-like empathy[6,7]. These kinds of overstated claim are even more evident when commercial enterprises have an interest in promoting the diagnostic value of a test.

Moreover, the translation of research from the laboratory to particular social practices, such as medicine or education, is not determined by the scientific innovation itself but by social and political decisions that are often shaped by many factors other than scientific evidence. These processes of translation, and the ethical dilemmas they raise, need to be analysed. And the analyses should not be viewed as merely the 'social work' adjunct to the 'hard science'. Research into the social and ethical processes of translation, and into the challenges that are often faced, should inform the work of researchers themselves and can help to ensure that this research does result in improvements in social practice. For example, challenges to the validity of psychiatric diagnoses have led to increasingly complex models of conditions such as attention deficit hyperactivity disorder (ADHD), as well as efforts to validate and standardize diagnoses for these conditions[8].

We have three main concerns about the potential use of psychiatric biomarkers. What is the best way to communicate the idea of a 'risk profile', and how might this affect personal identity? Given that human behaviour and psychiatric disorders arise from a complex set of factors, how can this complexity be retained when using information about biomarkers in the clinic and community? And what issues might arise from the commercialization of biomarkers, and how should they be addressed?

Personal identity and risk profiles

For children and adolescents, it has been common practice to identify those 'at risk' of psychiatric disorders, educational failure, or social and emotional difficulties, and to provide them with special resources as part of a social programme. However, identifying such children by using genetic and neurobiological biomarker information (potentially before symptoms develop) will be new. The lack of predictive power of biomarkers as risk factors for individuals, and the probabilistic nature of risk assessments based on them, raises specific issues. We are particularly concerned about biomarkers

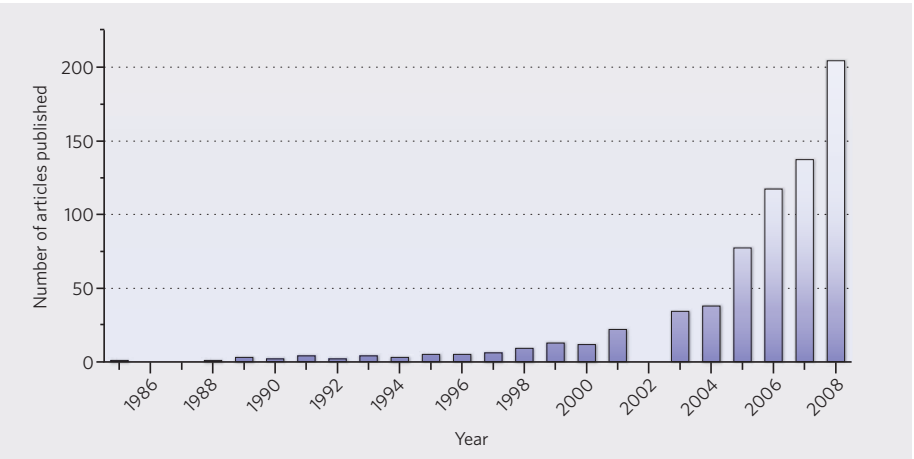


Figure 1 | Rising interest in biomarkers and psychiatry. In the past decade, psychiatric researchers, doctors and policymakers have become increasingly interested in finding objective biomarkers that will provide a more accurate and precise means of assessing actual and potential psychiatric conditions. This increased attention to biomarkers is reflected by the rise in the number of scientific articles on this topic. Data were obtained from a search of the ISI Web of Knowledge in May 2009 for articles with the term 'biomarker' and the word stem 'psy' in the topic field. No articles were indexed by ISI for the years 1986, 1987 and 2002. It should be noted that some early articles may have focused on non-psychiatric conditions and only tangentially addressed the psychiatric aspects. (Courtesy of J. Abi Rached, The London School of Economics and Political Science, UK.)

Box 1 | What are biomarkers?

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention[40].

Different markers can be used to make various assessments: to diagnose a condition; to predict the natural outcome for an individual with this condition; to predict whether the individual will benefit from a particular treatment and how aggressively to treat the individual; and to assess an individual's response to this treatment.

In a psychiatric context, biomarkers could be used to detect and assess, or predict the development of, not only psychiatric disorders but also personality or behavioural traits, and emotional or cognitive capacity. Biomarkers could also be used to inform treatment decisions. Examples of biomarkers are:

- specific variations in skin conductivity;
- specific patterns of neural activity in particular brain regions, detected by imaging techniques, such as functional magnetic resonance imaging (fMRI) and positron-emission tomography (PET);
- specific genetic sequences or single-nucleotide polymorphisms, identified by genetic screening;
- specific endophenotypes (intermediate traits in the chain of causality between genes and diseases), such as biochemical, neurophysiological or neuropsychological features.

that identify the risk of behavioural problems developing in children, as well as the risk of psychiatric diagnoses such as ADHD, conduct disorder and oppositional defiant disorder, all of which are directly or indirectly linked to delinquency, substance abuse, antisocial behaviour, personality disorder and/or criminality. Because antisocial behaviour by young people is a major social problem in many countries, if biomarkers for such behaviours are found to be present during early childhood screening, then children might be subject to intrusive medical interventions that focus on individual-level risk factors rather than on social and environmental risk factors[9]. Indeed, given the increasing use of psychotropic medication in very young children (3 to 5 years of age)[10], and the problem of inconsistently applied diagnostic thresholds in child psychiatry[11], it is possible that young children with a high risk profile for antisocial or criminal behaviour could be given drug treatments at a pre-symptomatic or subclinical stage.

The problems inherent in the probabilistic nature of risk assessment are not unique to psychiatry. But the consequences of psychiatric risk profiling for children and their families might differ qualitatively from those of profiling other medical conditions in childhood. At present, it is unclear what will happen when children are identified as being at risk of developing a psychiatric disorder or antisocial behaviour in societies that are suffused by anxiety about the adverse social consequences of such conditions. As risk profiling of children, using biomarkers, begins to replace efforts to identify biological or environmental causes, will ideas about the identity and the capacity of individuals begin to change? That is, how will people feel about themselves given their risk profile, and will others perceive them differently? Will 'risk' and 'potential' eventually dominate ideas of personal identity, health status and opportunity in rigid, coercive or stigmatizing ways? Will these ideas become institutionalized within education, law and policy? And how will such changes affect the life trajectories of children identified as at risk early in life?

There is a body of research on risk perception and stigma that addresses how people's judgements and beliefs about the genetic basis of psychiatric disorders affect their self-identity and their attitudes towards others with psychiatric diagnoses[12–14]. However, there has been little research on how these perceptions and attitudes might differ when they are shaped by information based on markers whose predictive value is probabilistic, although there are ongoing studies of these issues as they arise in the context of personal genomics[15]. Such research needs to be expanded to examine issues involved in the screening of individuals for the presence of psychiatric biomarkers, in order to evaluate the potential for children developing a negative 'risk identity'. It is also important to assess how biomarker information might reshape the beliefs, practices and decision-making of the people in a child's environment, including parents, teachers and health providers. And, given that young people themselves discuss these issues and share information about psychiatric conditions and interventions, it is important to understand further how biomarker information might affect both young people's identities as individuals and as members of various groups and their relationships with their peers.

Discoveries in neuroscience can, however, have a positive effect, particularly when they converge with patient activism, resulting in broader conceptions of self and possibility. For example, there are strong claims by some autism researchers and autism patient groups that biomarkers associated with autism indicate a divergence from the norm that is a source of creativity and special cognitive capacities[16]. In this case, how does biomarker information motivate individuals instead of inducing a fatalistic attitude? And how can such resilience be promoted in varied contexts, such as families and classrooms, when there is challenging news regarding a child's risk profile?

Another important issue is the potential consequences of claims about differences in risk prevalence between population groups. As is the case for physical diseases, evidence that a certain condition is more prevalent in one ethnic group than another is a double-edged sword. On the one hand, such evidence supports the development of strategies to tackle those disorders. On the other hand, it can lead to discrimination on biological or genetic grounds[17,18]. Although some people claim that identifying a high level of risk among ethnic minorities, single parents or the very poor can help to direct more attention and resources to the problems encountered by such groups, there is evidence that people in these groups are also more likely to be, and to feel, stigmatized by at-risk labels and psychiatric labels[19]. It is also important to determine whether children from these or other populations are more likely to be ascribed harmful risk identities than children from majority or advantaged groups when biomarkers are used to identify those at risk. In particular, genetic biomarkers that identify children from ethnic minorities as being at risk of problem behaviours and/or psychiatric conditions might build on existing assumptions about links between race, genetics and behaviour. And they might have a self-reinforcing effect, leading to efforts to undertake risk screens that genetically profile young people who have first been classified by race or ethnicity.

The only way to explore the 'social life' of biomarker information is to carry out a programme of detailed qualitative research in which all groups of individuals who are affected participate, including children. Such a programme of research should accompany current scientific research on biomarkers, even if these studies are not yet seeking translational outcomes. Because the expectations around biomarkers are so high, it is possible that any translational applications will be quickly implemented, without time for deliberation over the social and ethical issues. Prospective research on these issues is needed to inform policies and practices that will maximize the positive potential of biomarker information and protect individuals and families from harm.

Retaining complexity

Issues around communicating and explaining risks link directly to our second major concern: that the many factors contributing to behaviour and psychiatric conditions need to be considered in addition to information about biomarkers. The techniques used to measure biomarkers (mainly neuroimaging and genetic screening) produce images and numbers. These results seem to be precise and objective and therefore have great persuasive power, often greater than is warranted by their predictive power. Indeed, recent criticism over the level of funding of genome-wide association studies has highlighted their failure to uncover genetic markers that account for

most of the genetic contribution to the risk of common disease[20,21]. Similarly, magnetic resonance imaging (MRI) studies have uncovered much about the basic science of brain function and structure, but the predictive utility of brain-based biomarkers does not match the hype surrounding their discovery[22]. Although research on biomarkers is of considerable scientific interest and importance, the persuasive power of biomarkers will be greater than their clinical or social utility for the foreseeable future. It is important to prevent such persuasiveness from leading to reductionist explanations for complex behaviours or conditions in children. Information about a biomarker can help to build a risk profile for a particular condition or set of behaviours. But biomarkers alone, taken out of context of environmental influences, are unlikely ever to provide complete explanations for children's behaviour or a forecast of how children's lives will unfold. Biology is not destiny: biology provides information about potentials. So how can this level of complexity be retained when biomarkers move from 'the bench to the bedside'?

First, systematic assessments of the explanatory power of biomarkers for particular behavioural conditions and psychiatric diagnoses are needed. These assessments must focus on two areas. The accuracy and reliability of the techniques themselves — particularly genetic screening and neuroimaging — must be tested. And the validity of the findings must be assessed, given the problems with the methods and with study design. To take functional MRI (fMRI) studies as an example, studies involving children are often significantly underpowered[23]. Moreover, applying different statistical methods to fMRI data has been shown to deliver markedly different estimates of the significance of associations between brain-based biomarkers and cognitive–emotional traits[24]. There is also a lack of normative fMRI data for children, partly as a result of ethical concerns about scanning healthy children; therefore, children's scans are frequently analysed by comparing them with adult brain scans. Finally, outcomes of fMRI research are relevant at the population level but are not yet relevant for individual diagnoses and treatments. Published studies should be scrutinized for problems of validity, by meta-analyses for example (such as the recent critique of fMRI data reported in social neuroscience studies, which describes many of the results as 'voodoo' correlations[24]), and the results of such scrutiny should be reported and debated in scientific journals.

The next task will be to disseminate the results of these assessments in a comprehensive programme of public engagement. Dissemination strategies could take various forms, including media appearances, teacher and student education, publication in the popular press and artistic performances. Scientists must take up the challenge of collaboration with a variety of professionals from other disciplines, to build a better public understanding of behavioural conditions in children and the environmental and biological underpinnings of these conditions. Moreover, these activities should support the public's ability to think critically about current neuroscientific theories and the evidence that forms the basis of these theories, including the continuum of normal and abnormal behaviour, interactions between genes and the environment, causality and the direction of effects, and the probabilistic nature of genetic and neurobiological influences on behaviour and cognition. Given the public's lack of knowledge — misperceptions even — about genetics, neurobiology and behaviour[25–27], one of the most important components of this public-engagement programme will be to study its outcome: to examine which modes of dissemination and education are most effective at building an accurate public understanding in this area, and for which populations; and which initiatives are most likely to inspire sustained public engagement with these issues.

In addition, the translational activities of scientists and scientific teams who carry out research on biomarkers in children need to be investigated. How do efforts to find applications for biomarkers in the clinic, the classroom or the juvenile courtroom accommodate complex models of behaviours, models that are probabilistic, multidirectional and incorporate a variety of causes. And to what extent do proposed treatments or interventions either reinforce or undermine such models? For example, the rise in childhood diagnoses such as ADHD, bipolar disorder and social anxiety disorder suggests that the availability of effective drug treatments (such as stimulants, antipsychotics and antidepressants) undermines multicausal explanations of childhood behaviours. If biomarkers are used to inform strategies for treatment with psychotropic drugs, and if such treatment is used preventively, then there is a substantial risk that biomarkers will be seen as the primary or ultimate cause of behavioural conditions in a child.

Finally, biomarker-based family and educational interventions for children should be monitored by a research team that includes professional educators, mental-health professionals, ethicists and social scientists, working in collaboration with family representatives and teachers. Part of the remit of such research will be to identify the disadvantages and the advantages of policies that are informed by research on biomarkers, by evaluating the long-term educational, social and behavioural outcomes for particular groups of children. It will be especially important to carry out research that identifies and investigates the conditions under which genetic and neurobiological factors may come to define an individual's risk and potential without appropriate attention being paid to non-biological factors. At present, there is a strong emphasis on early identification and intervention programmes that assure children's long-term mental health and positive social contributions; therefore, a comprehensive system of monitoring the educational, social and ethical outcomes of such programmes for individuals and society is essential.

Commercial issues

Biomarkers are entering the public's awareness, through the activities of commercial enterprises such as 23andMe and other companies that offer genetic screening and brain scanning services directly to consumers. These ventures promise to provide individuals with significant information about their risk profile for numerous diseases and disorders, as well as information on personality and behavioural types. Clearly, businesses need to attract customers and will therefore emphasize what they consider to be the benefits of their services to potential consumers. This is the case not only for genetic tests but also for tests available on

Box 2 | Status of psychiatric biomarker research

Evidence of biomarkers associated with psychiatric disorders is emerging from studies in experimental neuroscience and psychiatric genetics, and many of these studies are improving our understanding of the neural pathways and mechanisms underlying such disorders. But, at present, only a few biomarkers show promise as robust predictors of psychiatric disorders. One prominent example is a variant of the monoamine oxidase A gene, which several studies have found to be associated with risk of antisocial behaviour, but only when there is a history of adversity in early childhood[41,42]. This discovery is the strongest indicator so far that biomarker information alone is unlikely to explain most of the variance in observed behaviours. Gene–environment interaction studies, rather than studies that search for single genetic polymorphisms, could deliver more robust results that have potential social and clinical applications.

In addition, researchers are identifying endophenotypes that can act as predictors of behavioural and psychiatric disorders. An example of a robust psychiatric endophenotype is abnormal eye-tracking movements in patients with schizophrenia[43].

In child psychiatry, a promising endophenotype is callous and unemotional traits[1]. Several studies have shown such traits to be strongly associated with the development of psychopathy in children. And although the neurobiological and genetic basis of these traits is not well understood, they are associated with a variety of biochemical, neuroendocrine and genetic markers.

At present, although neuropsychological testing is sometimes used to inform psychiatric diagnoses, biological information derived from brain scans or genetic screening is not yet used in clinical psychiatry.

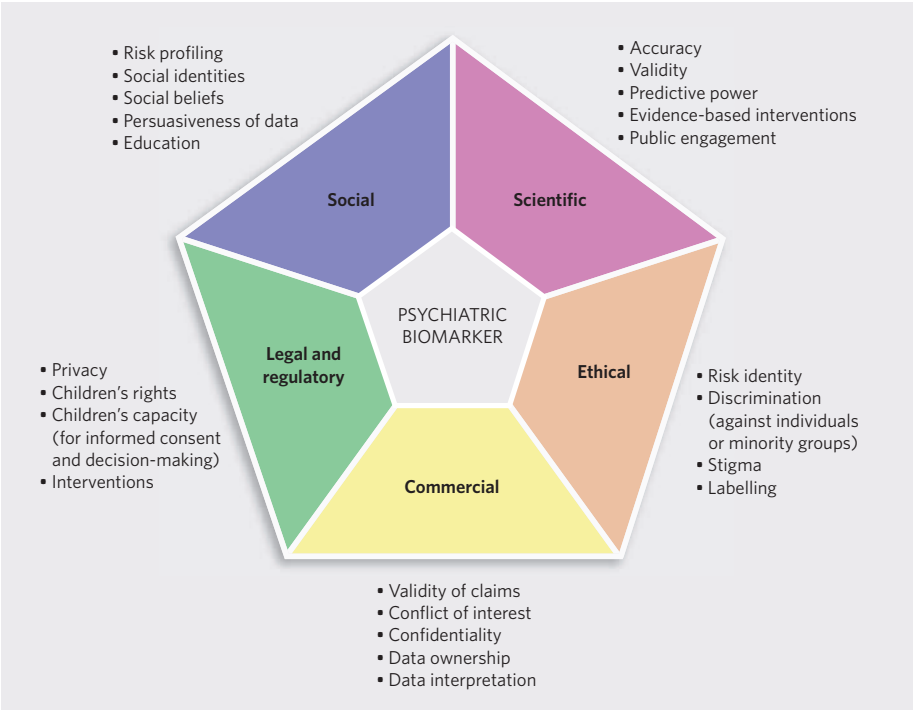


Figure 2 | Issues that influence the use of biomarkers in psychiatry and beyond. As the search for psychiatric biomarkers increasingly shapes research, as well as clinical and non-clinical settings such as classrooms and courtrooms, many issues need to be debated, studied and resolved. These include social, legal, ethical, commercial and scientific issues.

the Internet for everything from depression to osteoporosis and for home test kits such as those for measuring cholesterol levels and detecting signs of diabetes and bowel cancer. Many of these are a cause of concern to doctors and regulators, who consider that they provide information of dubious medical value and are sometimes dangerously misleading[28].

We are concerned that the commercialization and marketing of biomarker data about psychiatric conditions, personality traits, and emotional and cognitive capacities might have harmful consequences for the ways that families and children make vital decisions and view their future, and indeed how others make decisions on their behalf. The companies often argue that knowledge cannot be harmful and that they are making access to biomedical information more democratic[29]. They have also tended to resist regulation of their activities, branding this as paternalistic[30]. However, we think that social-science research is required to examine the extent to which such commercialization of biomedical information, and its availability on a direct-to-consumer basis, may encourage individuals and families — and perhaps medical and non-medical professionals — to develop unrealistic ideas about the explanatory power of this information. For example, evidence from the United States shows that patients who mention an antidepressant by name to their doctor are more likely to receive a prescription for that drug[31]. So will doctors be more likely to diagnose a child as having a particular psychiatric condition — or to provide medical treatments — if families come into the clinic on the basis of biomarker information that they have obtained themselves?

Little is known about how those who take advantage of commercial screening and scanning interpret and use information about personal biomarkers. Social research on the use of personal genetic risk information suggests that individuals do not always reshape their lives or identities around such information, even when the probability of developing a severe disease is high[32–34]. But will this be the case when the personal biomarker information is obtained by a child or by a parent acting on behalf of a child, and when the targeted biomarkers ostensibly reveal information that is closely associated with personal identity (such as personality traits, behavioural traits, and/or predisposition to certain thoughts and feelings)? Why would parents initiate commercial genetic screening or brain scanning for their child? Will biomarker information that provides a cognitive–behavioural risk profile for a young child affect not only how parents think about their child but also how they act on behalf of the child? In what circumstances might this have positive consequences for the child, and when might this information do harm? Much more information is needed about the social and behavioural consequences of the availability of personal biomarker information for children before evidence-based judgements can be made about the ethical issues raised by such technologies.

One key ethical issue is the capacity of children to understand the consent processes involved in submitting to commercial screens and scans, as well as their capacity to make decisions about the complex issues that results often raise. To what extent, and at what age, do children have the right to know about their personal biomarker profiles? Do children have the right to refuse to submit to genetic screens and brain scans when these are not clinically indicated? Conversely, should children have the right to submit their own samples to such companies or to choose to undergo brain scans to learn more about their risk of displaying certain behaviours or traits? Research into this issue is needed so that thresholds can be set for the age at which children are competent to provide consent for screening and scanning and to make informed decisions about the findings. The regulation of children's rights, capacity and consent in analogous contexts, such as birth control and cosmetic surgery, could inform thinking about children's access to personal biomarker information.

A related area of concern is confidentiality and ownership of information. Who has the right to manage a child's 'potential' when that potential seems to be made evident by a genetic screen or a brain scan? Medical lawyers and ethicists need to identify the appropriate precedents for ownership of such information: for example, is legislation around blood and tissue data held in biobanks relevant to personal data derived from genetic screening and brain scans, or is such information more analogous to medical records, access to which is covered by data protection legislation in the United Kingdom and privacy legislation about access to health-related data in the United States.

To understand children's capacities, and to evaluate the potential risks and benefits of the availability of psychiatric biomarker information to them (as well as to their families, their doctors and other authorities), empirical research must involve children themselves. Questionnaires and surveys alone cannot adequately capture children's thought processes or their capacity for making complex decisions. These approaches need to be supplemented by detailed interviews with children are required. Only then will there be enough evidence to underpin the development of appropriate regulation for commercial organizations, in order to maximize the benefits of access to biomarker information for children while effectively protecting them from harm.

Towards the future

So how might biomarkers be applied in the future? To take one example, in the United States and the United Kingdom, there are several high-profile research programmes that seek to identify biomarkers associated with risk of delinquency, antisocial behaviour or criminality in children[35–37]. Some of this research is being carried out with children identified as having behavioural problems such as hyperactivity and impulsiveness. In other studies, the research subjects are children diagnosed with conduct disorder, oppositional defiant disorder

or ADHD. Some researchers hope that such studies will establish biomarkers (for example, brain-based markers, cardiac markers or neuroendocrine system markers) that would allow at-risk children to be identified before they display serious behavioural problems. Children at high risk could then be treated before symptoms arise, to prevent the development of delinquent behaviours. For prevention to be successful in this high-risk group, interventions would need to take place very early, in the preschool years (ages 3 to 5)[38].

The aim of this research is to develop programmes of identification and intervention that will reduce the individual and social burdens of severe antisocial behaviour in adults. However, we are of the view that, at present, the assumption that better understanding of the neurobiological risk of delinquency will facilitate early identification and prevention efforts is unfounded. Certainly, carrying out early interventions on the basis of this assumption would be premature. For example, we would be very concerned if a 4-year-old boy displaying mild hyperactive behaviours were a candidate for treatment with stimulant drugs based on his hyperactivity being a predictor of future delinquency. Antisocial behaviour in adults arises from a complex course of neurological development in certain environmental and social contexts[39], and we do not think that there is enough evidence that biomarkers, at least as they are conceived and identified at present, provide a justifiable basis for intervention programmes of this type.

Moreover, the assumption that childhood diagnoses such as conduct disorder, oppositional defiant disorder and ADHD represent underlying biological impairments is problematic, given that these diagnoses have questionable validity, especially in the preschool population. Without ongoing social and ethical analysis, as well as careful thought by the researchers about their role in this process, the future use of psychiatric biomarkers could marginalize efforts to identify and address social and environmental factors associated with the development of antisocial and criminal behaviours in young people. It could also reinforce the use of problematic diagnoses and/or medical treatments to manage the current and anticipated behaviour of very young children. Such developments could lead to stigma and labels that affect children's psychological development, their social and educational opportunities, and their medical care and employment options.

Clearly, the aim of developing psychiatric biomarkers is to improve the well-being of children and of society in general. But given the many issues that need to be considered (Fig. 2), a comprehensive programme of research needs to be carried out before biomarkers can be used ethically and effectively in the clinic, courtroom, classroom and community. The programme that we have outlined here can be accomplished only through interdisciplinary interactions between neuroscience researchers, doctors, social scientists, ethicists, legal scholars, policymakers and those involved in commercializing biomarkers. We do not envisage one large research programme, however. Instead, we think that multiple smaller collaborations, built around a variety of clinical and non-clinical sites, will be most effective.

The involvement of social researchers and ethicists at such sites should not be ad hoc. Social, ethical and policy concerns should be integral to the design of each study. This will also help to ensure that multidisciplinary engagements are collaborative and constructive for all researchers involved. It is unclear whether mandating the inclusion of these concerns in all relevant scientific research programmes on biomarkers is an effective strategy. All genomics research programmes are required to include an ethical, legal and social issues component, and this has resulted in important research but has also created many problems, such as tensions among researchers from different disciplines, and a lack of integration of social and scientific research goals. It is probably more effective to identify which ethical issues are the most crucial and then to challenge the funders of such research programmes to encourage grant applicants (through higher ratings on proposals and direct feedback to applicants) to integrate ethical, legal and social issues research and the appropriate researchers into their grant proposals.

Non-scientific sites where collaborations need to take place include communities and schools that carry out early-intervention programmes for children, juvenile prisons and courtrooms, and corporations that commercialize biomarker information. The development of research programmes will depend on access to these sites. Again, funding mechanisms can be used to encourage collaborative engagements between people working at these sites and researchers. As research information is gathered across all sites, workshops and conferences can be organized to share knowledge and specifically to guide ethically informed, clinically relevant and socially effective policymaking around the use of psychiatric biomarkers in children. ■

Ilina Singh and Nikolas Rose are at the BIOS Centre, London School of Economics, Houghton Street, London WC2A 2AE, UK.
e-mail:i.a.singh@lse.ac.uk

1. Moffitt, T. E. *et al. J. Child Psychol. Psychiatry* **49,** 3–33 (2008).
2. Kirkwood, T., Bond, J., May, C., McKeith, I. & Teh, M. *Foresight Mental Capital and Wellbeing Project. Mental Capital Through Life: Future Challenges* (The Government Office for Science, London, 2008).
3. Goswami, U. *Psychol. Educ. Rev.* **29** (2), 2–8 (2005).
4. Maroney, T. A. The false promise of adolescent brain science in juvenile justice. The Gruter Institute Annual Conference (Squaw Valley, California, 18 May, 2009); at http://ssrn.com/abstract=1405367.
5. http://www.researchchannel.org/prog/displayevent.aspx?rid=4775
6. Brunner, H. G. *Ciba Found. Symp.* **194,** 155–167 (2007).
7. Dinstein, I., Thomas, C., Behrmann, M. & Heeger, D. *Curr. Biol.* **18,** R13–R18 (2008).
8. Swanson, J. *et al. Neuropsychol. Rev.* **17,** 39–59 (2007).
9. Romeo, R. *et al. Br. J. Psychiatry* **188,** 547–553 (2006).
10. Zito, J. M. *et al. J. Child Adolesc. Psychopharmacol.* **17,** 195–204 (2007).
11. Parens, E. & Johnston, J. *Child Adolesc. Psychiatry Ment. Health* **2**, 5 (2008).
12. Link, B. G. *et al. Psychiatr. Serv.* **52,** 1621–1626 (2001).
13. Phelan, J. C. *Trends Neurosci.* **25,** 430–431 (2002).
14. Phelan, J. *Am. J. Med. Genet.* **141B,** 688 (2006).
15. Prainsack, B. *et al. Nature* **456,** 34–35 (2008).
16. Silverman, C. *Biosocieties* **3**, 325–341 (2008).
17. Collins, F. S. *Nature Genet.* **36,** S13–S15 (2004).
18. Rose, N. *Biosocieties* **3,** 423–439 (2008).
19. Thornicroft, G. *Shunned: Discrimination Against People with Mental Illness* (Oxford Univ. Press, 2006).
20. Goldstein, D. B. *N. Engl. J. Med.* **17,** 1696–1698 (2009).
21. http://scienceblogs.com/geneticfuture/2009/04/personal_genomics_is_not_dead.php
22. Weisberg, D. S. *et al. J. Cogn. Neurosci.* **20,** 470–477 (2008).
23. Seyffert, M. & Silva, R. *Curr. Pediatr. Rev.* **1,** 17–24 (2005).
24. Vul, E., Harris, C., Winkielman, P. & Pashler. H. *Perspect. Psychol. Sci.* **4,** 274–290 (2009).
25. Green, M. J. & Botkin, J. R. *Ann. Intern. Med.* **138,** 571–575 (2003).
26. Schick, A. *Am. J. Bioeth.* **5,** 36–38; W3–W4 (2005).
27. Foster, M. W. *et al. J. Med. Ethics* **32,** 635–638 (2006).
28. Hudson, K., Javitt, G., Burke, W., Byers, P. & ASHG Social Issues Committee. *Am. J. Hum. Genet.* **81,** 635–637 (2007).
29. http://www.23andme.com/about/press/20080909b
30. Foster, M. W. & Sharp, R. R. *Per. Med.* **5,** 399–404 (2008).
31. Kravitz, R. L. *et al. J. Am. Med. Assoc.* **293,** 1995–2002 (2005).
32. Huniche, L.*Outlines Crit. Soc. Studies* **3,** 35–52 (2001).
33. Lock, M. *Curr. Anthropol.* **46,** S47–S70 (2005).
34. Lock, M. *et al. Public Underst. Sci.* **15,** 277–300 (2006).
35. Viding, E., Blair, J. R., Moffitt, T. E. & Plomin, R. *J. Child Psychol. Psychiatry* **46,** 592–597 (2004).
36. Decety, K. J., Michalska, Y. A. & Lahey, B. B. *Biol. Psychol.* **80,** 203–211 (2008).
37. Beauchaine, T., Neuhaus, E., Brenner, S. & Gatzke-Kopp, L. *Dev. Psychopathol.* **20**, 745–774 (2008).
38. Beauchaine, T. *Biomark. Med.* **3,** 1–3 (2009).
39. Blair, R. J. R. *Brain Cogn.* **55,** 198–208 (2004).
40. Biomarkers Definition Working Group. *Clin. Pharmacol. Ther.* **69,** 89–95 (2001).
41. Caspi, A. *et al. Science* **297,** 851–854 (2002).
42. Foley, D. L. *et al. Arch. Gen. Psychiatry* **61,** 738–744 (2004).
43. Holtzman, P. *Brain Res. Rev.* **31,** 350–356 (2000).

Toxicology for the twenty-first century

Thomas Hartung

The testing of substances for adverse effects on humans and the environment needs a radical overhaul if we are to meet the challenges of ensuring health and safety.

Synthetic chemicals have been components of consumer products for just over a century. A system for identifying which chemicals pose a danger to individuals and the environment was first put in place about 80 years ago. But after several productive decades, in which a patchwork of testing approaches was formed, fewer and fewer of the latest scientific developments were incorporated. The system of regulatory toxicology fell asleep, much like the fairy-tale character Snow White when she bit into the poisonous apple. In the case of toxicology, the poison was international guidelines. This international harmonization was tempting because it allowed manufacturers and suppliers to use fewer resources, and it overcame barriers to trade in global markets. But implementing these guidelines came at a price: the slow and complicated international consensus process hindered self-criticism and modernization of the field of toxicology.

There is almost no other scientific field in which the core experimental protocols have remained nearly unchanged for more than 40 years. Yet consumers continually increase their expectations about the safety of products. One recent effect of this was the instigation of the largest safety assessment of chemicals that has ever been carried out: the European Union introduced the regulation known as Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) by legislation in 2007. Whereas new chemicals have been systematically evaluated in the European Union and the United States for about a quarter of a century, the safety of any chemicals produced before 1981 (which includes 97% of the major chemicals in use, and more than 99% of chemicals produced by volume) has not necessarily been properly addressed. In fact, it is estimated that data for 86% of the chemicals are lacking, and the REACH process seeks to redress this. The regulation affects 27,000 companies, which are required to provide information on the toxic properties and uses of 30,000 chemicals, after a pre-registration phase in 2008. But REACH might turn out to be like the prince whose kiss awoke Snow White after a long sleep, rousing toxicology at last.

Defining the problem

So what is wrong with the current approach to toxicology testing? An ideal study to understand whether an agent is harmful to humans would require an extremely large number of human subjects who are representative of the diversity of humans and who are unknowingly exposed to the agent under realistic conditions. All possible effects should then be assessed. If there is any deviation from these experimental conditions, which are unrealistic and unethical, the study will provide only an approximation of the real situation — it is a model. The crucial question therefore is how useful are the current models, which are mostly animal models, and how incorrect are they? Given that about €10 billion (US$14 billion) is spent on animal experimentation worldwide every year (about €2 billion of which is for toxicological studies), and given that more than 100 million experimental animals are used[1] and that products worth €5.6 trillion are regulated by such testing, the question is certainly appropriate. It encompasses four main issues.

The first issue is the extent to which animal models reflect human responses. It is clear that the use of animals has limitations[2]: we are not 70 kg rats; we take up substances differently; we metabolize them differently; we live longer (allowing certain diseases to develop and prompting evolutionary adaptations to protect against them); and we are exposed to a multitude of environmental factors. However, few studies have systematically measured the accuracy of animal models. In one example, results from animal models were compared with information from poison centres: comparing the dose of a chemical that is lethal to 50% (LD_{50}) of rats tested and the lethal concentration of the same chemical in the blood of humans showed a rather poor correlation (coefficient of correlation of 0.56; unpublished observations from an international validation study[3]). Similarly, in another study, 40% of the chemicals that irritated the skin of rabbits were found not to be irritants in the skin 'patch test' in humans[4].

Given the overall lack of data, this problem can be considered in more general terms by looking at how one species models for another. In several animal species, similar experiments with the same agents have been carried out, and there is no reason to assume that, for example, mice, rats and rabbits predict each other's response to a lesser extent than they predict that of humans. Typical results from such studies show agreement between animal species for 53–60% of chemicals[5,6].

Similar results have also been obtained for pharmaceuticals (as opposed to chemicals) that have been tested in humans. In one study, 43% of toxic effects in humans were correctly predicted by tests in rodents, and 63% by tests when non-rodent animals were also included[7]. It is clear therefore that many adverse effects are not uncovered by such traditional tests. This is also evident in data from the pharmaceutical industry, showing that 20% of the failure of drug candidates occurs as a result of toxicity only after the drugs have been administered to humans in clinical trials[8]. And it is estimated that 6.7% of hospitalized patients experience unexpected adverse reactions to drugs (1 in 20 of which are fatal)[9], showing the limitations of anticipating toxic effects from preclinical animal studies. To improve the toxicity assessment, tests are often carried out in two animal species: usually substances that show no toxic effect in one species are then tested in another species to improve the likelihood of finding any toxic properties. This increases the sensitivity of testing (that is, it increases the proportion of toxic substances that are found) but at the cost of increasing the number of false positives (when non-toxic chemicals seem to be toxic in the tests carried out).

The second key issue facing animal testing relates to the study design, particularly to the highly precautionary (conservative) approach that is taken at present. To limit costs and animal numbers, toxicity testing is typically carried out with the maximum dose of the chemical that can be tolerated, which has previously been determined. Such doses can be more than 1,000-fold higher than the doses intended for humans (in terms of milligrams per kilogram body weight, for example). This strategy yields many false positives and further diminishes the correlation between findings in animal models and humans[10].

The third issue is the testing of multiple end points, which also contributes to false-positive results. When enough end points are studied, positive associations will always be found. This is elegantly illustrated by a study that searched for correlations between disease and zodiac sign in the health records of 10 million residents of Ontario, Canada[11]: those born under the sign of Leo had a significantly higher probability of developing a gastric haemorrhage than individuals of other zodiac signs, and Sagittarians had far more fractures of the humerus over the period analysed. The explanation for this is simple: a total of 223 medical conditions were studied in a single population, and examining so many variables inevitably results in some extreme clustering of random results. Similarly, in toxicological studies, a large number of end points are measured: about 40 in repeat-dose toxicity studies (long-term studies in which animals are exposed to a chemical for a month to a year, and the effects on many organs are studied); and 80 in reproductive toxicity studies (in which adverse effects on the reproductive system, from fertility to embryonic malformations, are analysed). Unavoidably, some end points will be positive, and the group sizes used are too small to allow statistical correction for this.

Given that the cost of current tests is several hundred thousand euros per substance, large increases in group sizes are unrealistic, in addition to this being undesirable from the perspective of animal welfare. Therefore, all positive results have to be recorded as true positive results. However, this is less undesirable in the risk-assessment process than one might think, because the positive findings are simply used to establish the 'lowest observed effect level' (that is, the smallest amount of a substance that causes an observable change in the organism being studied). But, because the maximum tolerated dose is being used, there is usually a large safety margin (typically a factor of 100), so the substance could still be used even if seems to be toxic at high doses (with the exception of chemicals observed to have tumour-inducing properties, as these effects are generally thought to be relevant at much lower doses). It is thus often not important whether a positive result is an artefact. It is relevant, however, to those who later need to reproduce the presumed organ toxicity to validate an alternative approach, because false-positive results are difficult to reproduce when a different test is used. In addition, whether or not a positive result is false is unlikely to be noticed, because most regulatory tests are carried out only once[12] and because toxicological studies are often not reported publicly[13]. So the self-corrective mechanisms of science are not in place: there is no cross-referencing between similar studies in different laboratories.

The fourth issue concerns the prevalence of chemical effects on health. In other words, how many chemicals actually have hazardous properties[14]? Despite the use of highly precautionary tests, more than 87% of chemicals registered as new chemicals over the last 25 years are not acutely toxic in current tests; 93% of them do not irritate the skin[15]; and only 2–3% impair the reproductive cycle[16].

So toxicological studies search for a rare hazard with imperfect models. What are the consequences of this?

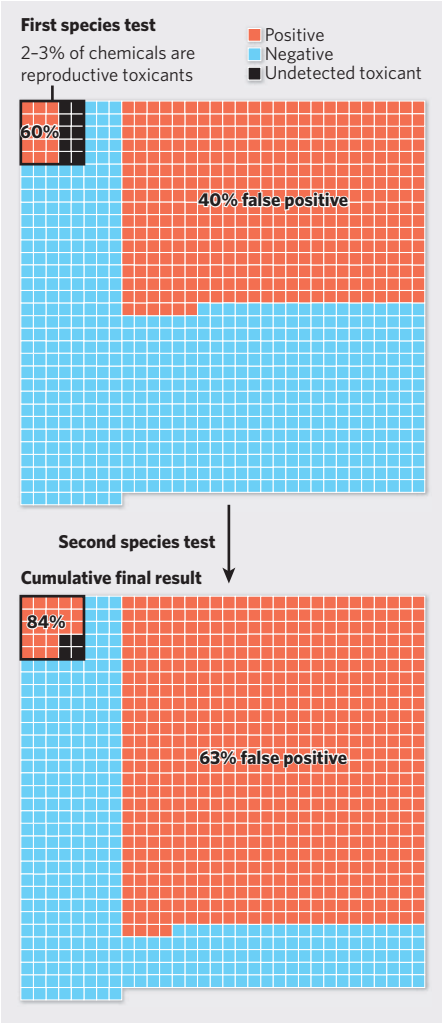


Figure 1 | The consequences of searching for rare hazards using imperfect tests. For reproductive toxicity testing, the concordance between animal species is about 60%. So roughly 40% of non-toxic chemicals will yield false-positive results. This problem is compounded by the standard practice of testing these chemicals that yield negative results in a second species to increase the number of hazardous substances identified.

False-positive issue

Take, for example, reproductive toxicity testing under the REACH legislation. All chemicals that were marketed before 1981 and are produced at more than 100 tonnes per year in the European Union will be subject to testing: about 5,500 of the 30,000 chemicals covered by REACH. It is estimated[16] that about 2.5% of these (138 substances) are true reproductive toxicants in humans (Fig. 1), and the goal of toxicological testing is to identify these. The reported concordance between species is about 60% for reproductive toxicity testing, using the two-generation study in rats (in which toxic effects are followed not only in the offspring of exposed rats but also, after further mating, in the next generation). Between animals and humans, however, this concordance might be even lower, owing to the high-dose, precautionary approach. So, when testing 5,500 chemicals with a test that is 60% accurate, 83 of the 138 reproductive toxicant will be found, but about 2,145 substances (almost 40%) will yield a false-positive result. The standard procedure would then be to test the apparently non-toxic substances in another animal species. Given the same accuracy, in rabbits or mice, 40% of the 3,272 substances that showed negative results in the first test (1,309 chemicals) will test as false positives. At the same time, 60% of the 55 true toxicants (33 chemicals) that were missed in the first test, in rats, will be found.

In total, 116 of the 138 true reproductive toxicants (84%) will be found, and 3,454 non-toxic chemicals will be found to be toxic (a total of 63% false-positive findings). These results might therefore restrict the use of a large number of these substances, which are subject to testing because they are produced in the highest quantities in Europe[17]. This scenario might be difficult to believe, but an analysis of reproductive toxicity studies for chemicals between 1981 and 2007 confirms this[16]: in 27 years, 72 chemicals reached a production volume that triggered reproductive toxicity tests. Of these, 41 (57%) tested positive, as the above calculation (of 63%) would suggest.

There are several caveats though. The above scenario might be too pessimistic because the correlations between species are biased by the inclusion of more chemicals that test positive in at least one species (because, in the past, a second test was often carried out to challenge the result). In addition, triggers others than production volume might have indicated the need for testing a substance: that is, if substances are tested because they are suspected of being toxic to the reproductive cycle, then this biases the number of positive results in the database. Nonetheless, it is unlikely that we can afford to falsely assign a large proportion of high-production-volume chemicals as reproductive toxicants. This will unnecessarily restrict the use of many substances, require large and expensive efforts to replace chemicals that are widely used, and create unnecessary fears in consumers about previous exposure. It might also prompt a situation similar to that for pharmaceuticals: if such results are obtained for a drug that is in the late stages of development (when it is already certain that the drug has financial value), then large amounts of toxicological work are required to determine whether the animal studies are in any way relevant to humans so that a valuable substance can be saved.

Another important issue is that the tests for each chemical require an average of 3,200 animals for a single two-generation test[17] — a total of 17.6 million animals for 5,500 substances — and the current REACH testing guidance

for industry does not include much scope for waivers or alternatives. Even if the use of alternatives to animal studies, such as cell-culture-based testing, were feasible, such methods do not have fewer limitations[18], except for ethical ones. And particularly in the field of reproductive toxicity, alternative methods are only being developed[19], and the cost–benefit ratio of using these for large-scale screening programmes still needs to be established.

Towards a solution

It is unlikely that researchers will suddenly produce new tools and design new methods with great accuracy. The solution to using fewer animals and making better predictions in the mid-term is to design integrated testing strategies. At present, the typical process is to use a default animal test and then, in some cases, to use cell-culture and computer-based methods to define the mode of action of the toxin and to interpret and balance the results further. But the best opportunity to improve regulatory toxicology lies in strategies in which optimal use is first made of all existing information about a substance and structurally similar substances, and then information is gained by approaches that do not involve animal testing, leading to targeted animal testing only if necessary. Such strategies will ideally include decision points that depend on interim results. An example of such a strategy is shown in Fig. 2.

The simplest testing strategy would combine two different approaches, such as a screening approach (a method to identify 'suspicious' substances with less effort and allowing false-positive results) and a confirmatory one (which may be more sophisticated and specifically identifies hazards with higher certainty). All substances that test positive during the screening approach or another prioritization step would enter the confirmatory stage, which would consist of, for example, a battery of mechanistic tests examining relevant pathways of toxicity. Instead of testing a large number of substances that includes few true toxicants by using one definitive test, this new approach would increase the number of true positives entering the confirmatory stage by creating a subset of suspicious substances, offering more evidence about whether a chemical is hazardous than the screening test alone. Alternatively, analysing which end points (for example, which of the up to 80 end points measured in a reproductive toxicity study) actually lead to classification as toxic or non-toxic in a particular animal test might allow researchers to identify the end points for which dedicated tests are required[16].

Despite the advantages of such a change in approach, several difficulties are apparent. It would first require acknowledging and analysing the limitations of the current approach. One central problem here is that the current system is convenient for the key players: namely, the regulators and the regulated industry. At present, a clearly defined set of tests, which make predictable demands on time and costs, is carried out. Then, the books are closed, and industries' liability is minimized. Every integrated testing strategy with decision points in its course will bring this simple procedure to an end, making the uncertainties more evident, as well as the fact that only the probability of a particular hazard is assessed. Any shortcomings identified in current practices will mean that products need to be examined further and will open up liabilities again.

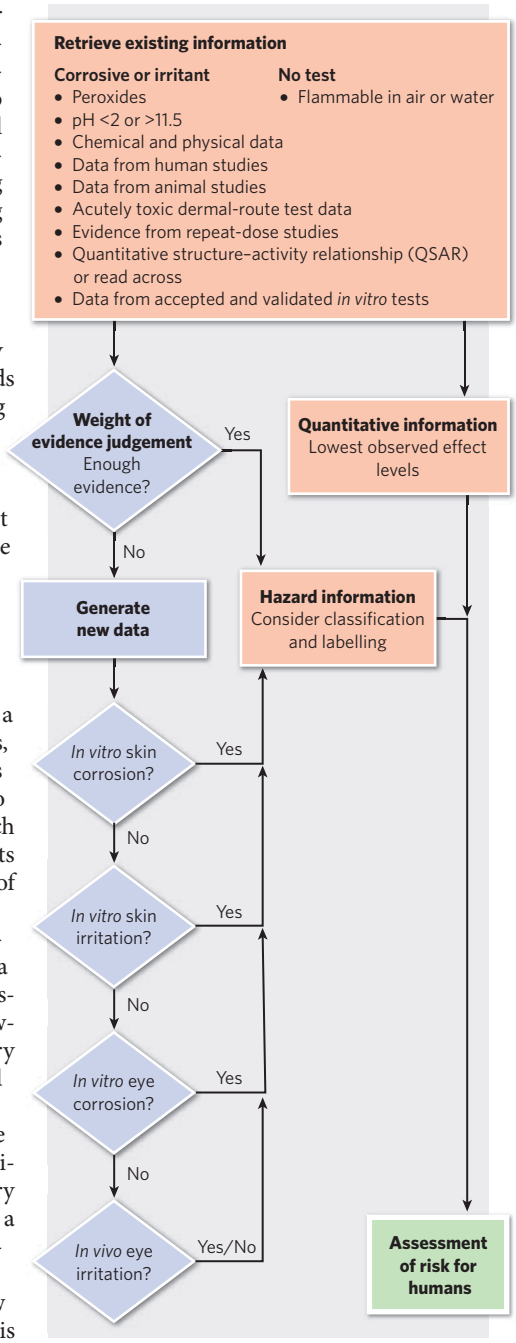


Figure 2 | Integrated testing strategy for eye and skin toxicity. This strategy from the REACH guidelines for industry is one of the first examples of an integrated testing strategy. The sequence includes decision points and involves assessing the existing information and then carrying out various *in vitro* tests, with animal tests being used only as a last resort (*in vitro* tests for eye irritation are being validated at present).

This situation is similar to that for pharmaceuticals. In fact, the current risk-assessment methodologies for chemicals are derived from those for preclinical studies of pharmaceuticals. However, for pharmaceuticals, there are two further steps in the process: clinical trials in humans and post-marketing surveillance (in which data are collected after a drug has been released onto the market). A considerable proportion of drug candidates (8–30%) fail because of safety problems in humans[20], despite having passed the entire toxicological programme of animal testing. Many of these safety issues are minor, for example nausea or a transient increase in the concentration of liver enzymes, but major chronic effects are not assessed at this stage. In addition, biologically active substances such as drugs often produce side effects as a result of their intended actions on human physiology (an effect known as 'excess pharmacology'); this is less of a problem for other areas of chemical use, in which the chemicals are not usually intended to affect the human body. But even though drugs undergo additional trials in human volunteers and patients, in my opinion there is always a need to follow up products after marketing, as illustrated by the anti-inflammatory drug Vioxx. Similarly, the possible hazards of chemicals in consumer products will probably need to be followed up more intensively after marketing.

Today, the pharmaceutical field is again driving changes in safety testing. With human proteins or antibodies (collectively known as biologicals) making up about half of the new drugs entering the market, classical toxicology is largely useless, because these proteins mostly have species-specific actions and animals raise antibodies to them, limiting the value of animal testing. This has created pressure to develop human-cell-based models for these biologicals, and other areas of toxicology will benefit from this. The inadequacy of current methods is also evident for new products such as genetically modified food and animal feed[21], functional food (food with intended health effects), and nanoparticles[22], creating an additional demand for new testing methods. Similarly, current methods are not tailored to assess the risk of acute poisonings associated with chemical accidents, or biological or chemical weapons[23].

REACH will also be a key instigator of change. This is partly because unexpected positive test results for important chemicals will trigger a review of the approaches — it is unlikely that important chemicals with decades of use will be abandoned easily, without raising doubts about the assessment. In addition, the legislation itself already represents a revolution in safety-assessment practices. Over the past three decades, internationally agreed (animal) testing guidelines have set out precisely how data must

be obtained, whereas REACH calls for the integrated use of all methodologies and for the use of animals as a last resort (with certain obstacles in place). So REACH calls for more flexibility and for tailored approaches. In terms of REACH, the test guidance for industry that has been developed in the past three years guides scientists through the combined use of existing data, and *in silico* (computer-based), *in vitro* and *in vivo* approaches. The greatest challenge will be to standardize these approaches in test guidelines and to reach international agreement on them. It is reasonable to assume that at least five times more guidelines will be necessary to accommodate the new approaches, an enormous challenge to the regulatory community.

But the challenge goes one step further: for each new method, test guidelines need to be not only agreed but also implemented. An interesting test case is the local lymph-node assay, which is used to predict whether topical application of a chemical to the skin will induce an allergic response. In 2002, the assay was internationally agreed by the Organisation for Economic Co-operation and Development (OECD) as the preferred animal model for studying skin allergies, but it has been seldom used until recently. Since 2002, less than 10% of new chemicals have been tested in this way, as indicated by notifications to European regulatory bodies. Applying a new method is hindered by, on the one hand, tradition and established practices and, on the other hand, obstacles such as the absence of international agreements with countries in important economic markets (for example, Brazil, Russia and China have not yet necessarily accepted the new OECD approaches).

International companies tend to use the traditional test until the last important market has accepted the new approach. So the banning of the original test method when alternatives become available is the prime opportunity to force a change. The OECD have only banned one test so far, however: the classical LD_{50} test, which required 45 rats for testing each substance, was abandoned in 2000, when three validated alternatives were introduced, requiring only 8 to 15 animals to test one substance. In other cases, the traditional animal tests have not been banned or modified when alternatives were introduced, so the original tests can still be carried out for regulatory purposes if justification is provided. But when a new approach does not suit all needs (that is, it is not appropriate for all chemicals or accepted by all member states), it is difficult to remove the traditional guidelines. The regulators must then urge that the new approach be used, to reinforce its implementation. For this to work, the advantages of the new test or the shortcomings of the old test need to be made evident, and to be credible, this assessment must have a sound and objective basis. The problem is that established practices have become intertwined with scientific insights during the decades in which toxicological tests have been shaped, and political compromises around such tests have been made.

Clinical medicine has a similar problem in that diagnostic and therapeutic approaches need to be objectively appraised so that the best decision can be made for each patient. Here too, new scientific approaches are interwoven with traditions, financial compromises in terms of health care, and so on. In the past couple of decades, the most important development in this area has been the evidence-based health-care movement, steered by the Cochrane collaboration[24]. Using structured reviews, consensus processes and meta-analyses, a series of 5,000 guidance documents has been developed. These provide the best available consolidation of the evidence in a particular field.

It is tempting to translate this evidence-based approach to toxicology[25], and a similar movement has been initiated. A realistic assessment of the methods used in toxicological studies will help to improve these tools and to integrate them into testing strategies. At the same time, it will be important to find ways to combine information from various studies, both systematically and quantitatively. The difficulties entailed are illustrated by the results of 29 independent risk assessments of the industrial solvent trichloroethylene: 6 studies deemed it non-carcinogenic; 10 found it to be carcinogenic in animals but unlikely to be carcinogenic in humans; 9 found it a plausible carcinogen in humans but with negative epidemiological findings; and 4 found it a plausible carcinogen in humans, with positive epidemiology[26].

Future visions

So it is clear that the current system of testing needs to change. Moreover, the individual testing tools have limitations and are inadequate for toxicology in the twenty-first century. To resolve this, I propose a three-step solution (Fig. 3). First, the limitations of the current tools need to be objectively assessed, and a better understanding of their uses is needed (for example, we need to analyse the prevalence of particular hazards because appropriate test strategies depend strongly on whether the hazard is rare or frequent). Second, in the mid-term, the various approaches need to be integrated into testing strategies, making the best use of the existing methods by combining them strategically. And, third, an entirely new system is urgently needed and should be built from scratch, using modern methods.

The basis for such a new system has emerged over the past two decades: advances in cell-culture techniques have enabled biological phenomena to be studied *in vitro*, unlike when toxicological experiments were first designed. In fact, most data generated in the life sciences now originate from studies of *in vitro* systems. This change in experimental approach required not only the accumulation of experience in these new techniques but also the provision of standardized equipment, materials and training. Early cell-culture-based experiments were

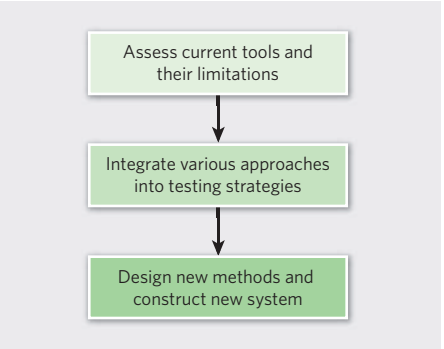


Figure 3 | Towards a new toxicology. The toxicology community needs to take three main steps to arrive at a new system of toxicology.

relatively simple, but they evolved rapidly, with many researchers now using three-dimensional ('organotypic') cultures that resemble organs in structure and function. Even one of the last big challenges in cell culture — the lack of availability of primary human cells (usually only sourced from surgically removed tissues with the notable exception of blood cells) — is now increasingly being overcome by isolating or generating human stem cells, from which most of the cell types in the body can be produced[27,28].

The avenue now opening for designing a new regulatory toxicology originates from the combination of bioinformatics and biotechnological approaches that yield huge amounts of information[29,30]. Three important technologies developed during the past decade have entered the field of toxicology[31,32]: 'omics' technologies (such as genomic and proteomic analyses), imaging techniques and robotized testing platforms. The testing platforms allow high throughput of samples, enabling large numbers of substances to be tested under standardized conditions. Omics technologies and imaging methods compile enormous sets of information about a single compound. Together, the three technologies not only allow researchers to 'fish' for new biological markers of specific toxic effects but also increasingly allow the deduction of patterns (or signatures) that are characteristic of certain toxic effects. By also harnessing advances in bioinformatics and *in silico* modelling, this information can be mined and then integrated with knowledge from other areas of the life sciences[33]. Such integration of information will be particularly important for investigating cellular pathways and should allow the cross-fertilization of ideas between toxicology and basic science[34]. The combination of biochemical knowledge of cellular pathways with genomics, proteomics and metabonomics (the study of metabolic responses to environmental factors, drugs and diseases) is already advancing as systems biology, and systems toxicology is a new subbranch of this field.

Such a systems approach was put forward as a toxicology for the twenty-first century in a 2007 report by the US National Academy of Sciences on behalf of the Environmental Protection

TABLE 1 | TOWARDS A NEW REGULATORY TOXICOLOGY

Scientific developments	Strategic developments
Mapping of pathways of toxicity by combining 'omics' technologies and data mining	Objective assessment of current practices (evidence-based toxicology)
Organotypic cell cultures and human tissues derived from stem cells	Guidance on Good Cell Culture Practice (an initiative for standardizing practices globally), Good Modelling Practices
Modelling of kinetics of substances (especially physiologically based pharmacokinetic modelling) in an organism for extrapolating from effective tissue concentrations to whole-organism doses	Systematic composition of testing strategies (mainly decision theory and sensitivity analysis)
In silico methods such as quantitative structure–activity relationship (QSAR) modelling	Validation of complex methodologies, in the absence of a gold standard
Imaging technologies and automated testing	Change management based on cost–benefit analysis
Integration of technologies	More communication

Agency (EPA)[35]. And this has already led to the formation of a coalition of US agencies[36] and to a revised toxicity testing strategy by the EPA[37], with the EPA's ToxCast programme[38] being closest in terms of vision to the new process required. So, in Europe, the prince who awakened toxicology was politics (with REACH calling for a new safety testing approach on a large scale, which was assisted by the animal testing ban in the seventh amendment of the cosmetics directive)[39]. By contrast, in the United States, it was science responding to the EPA's request for a new vision. What lies ahead, however, must be an entirely scientific process. Furthermore, the political process is necessary to make the funding available so that the political will can be put into practice. The dimensions of the project call for a global programme. The groundwork for such an effort has been laid by increasing awareness of the shortcomings of current methods, as well as emerging technological opportunities and political demands. The opportunity to create a new regulatory toxicology lies in a programme, similar to the Human Genome Project, that analyses the interactions of small molecules with cells. Such a programme will provide the molecular biological tools to switch cellular pathways on and off and to identify 'druggable' targets, and it will uncover the cellular pathways of toxicity, knowledge that is needed for a new way of approaching toxicology.

The main challenge is to design a new system of regulatory toxicology. Toxicology has grown step by step over a century to adapt to increasing and changing demands. Instead of amending the current patched-together system, a new system must be designed. And what is constructed from scratch with today's understanding and technologies will differ markedly from the current regulatory toxicology. We must forgo the approach that has been taken so far, which has been to add a new piece or to replace an old piece of the system: for example, by validating new tests that are each designed to substitute for a particular toxicological tool. This might deliver new ways to handle biologicals, nanoparticles and so on, but it will not solve the inherent shortcomings of an outdated architecture.

The necessary science seems to be available, but are the necessary scientists also available? Regulatory toxicology has not been the most appealing research area in the past. It could hardly compete for the best students with areas such as molecular biology, immunology and stem-cell biology, which have been advancing rapidly. However, science is invigorated when there are sufficient challenges and funding. And there is money available: I recently estimated that, in Europe, fulfilling toxicological regulations costs about €600 million per year for products that are traded at €1.7 trillion[40]. Generating data to comply with the REACH legislation will cost €8.8 billion using today's tests[40]. This is stimulus for a large number of targeted developments so that the process becomes better, quicker and cheaper, if scientists were only aware of this.

The scientific challenge laid down by this new vision of toxicology should appeal to scientists and to the commercial providers of solutions, mostly small-to-medium enterprises that are involved in commercializing new biotechnologies. Some key areas are listed in Table 1. But the challenge itself will not result in a new regulatory system. It will be important to open up regulators to the possibilities of a new system so that they give up on the old system and do not just use the new system as another patch for the old one (or as "useful additional information", as a regulator would say). For this revolutionary change to occur[41], the shortcomings of current methods need to be mapped and considered, and the transition from the old to the new approach needs to be steered[42]. This process needs to include standardization, validation and quality assurance of the new approaches, as well as the systematic integration of these approaches into testing strategies. There is a profusion of new concepts and technologies at present, but what is lacking is communication between stakeholders. Giving direction to the current stakeholders and to those in neighbouring disciplines who are not yet aware of the emerging opportunities, and allowing synergies to develop between approaches, might be even more important than the individual technological developments that are required. Promoting this process could be the real challenge for toxicology today. ■

Thomas Hartung is in the Department of Environmental Health Sciences at the Johns Hopkins University Bloomberg School of Public Health, and is Doerenkamp-Zbinden Professor and Chair for Evidence-based Toxicology. He is director of the Johns Hopkins Center for Alternatives to Animal Testing (CAAT), Baltimore, Maryland, as well as professor of pharmacology and toxicology at University of Konstanz, Germany.
e-mail: thartung@jhsph.edu

1. Taylor, K., Gordon,N., Langley, G. & Higgins, W. *Altern. Lab. Anim.* **36,** 327–342 (2008).
2. Hartung, T. *ALTEX* **25,** 3–9 (2008).
3. http://iccvam.niehs.nih.gov/docs/acutetox_docs/ATpanelrpt06/ ATpanelrpt.pdf
4. Basketter, D. A., York, M., McFadden, J. P. & Robinson, M. K. *Contact Dermatitis* **51,** 1–4 (2004).
5. Gottmann, E., Kramer, S., Pfahringer, B. & Helma, C. *Environ. Health Perspect.* **109,** 509–514 (2001).
6. Schardein, J. L., Schwetz, B. A. & Kenel, M. F. *Environ. Health Perspect.* **61,** 55–67 (1985).
7. Olson, H. et al. Regul. Toxicol. Pharmacol. **32,** 56–67 (2000).
8. Stevens, J. L. *Chem. Res. Toxicol.* **19,** 1393–1401 (2006).
9. Lazarou, J. *J. Am. Med. Assoc.* **279,** 1200–1205 (1998).
10. Rietjens, I. M. C. M. & Alink, G. M. *Chem. Res. Toxicol.* **19,** 977–981 (2006).
11. Austin, P. C., Mamdani, M. M., Juurlink, D. N. & Hux, J. E. *J. Clin. Epidemiol.* **59,** 964–969 (2006).
12. Purchase, I. F. H. *Toxicology* **202,** 1–20 (2004).
13. Wandall, B., Hansson, S. O. & Rudén, C. *Arch. Toxicol.* **81,** 605–617 (2007).
14. Hoffmann, S. & Hartung, T. *Toxicol. Sci.* **85,** 422–428 (2005).
15. Hoffmann, S., Cole, T. & Hartung, T. *Regul. Toxicol. Pharmacol.* **41,** 159–166 (2005).
16. Bremer, S., Pellizzer, C., Hoffmann, S., Seidle, T. & Hartung, T. *Curr. Pharm. Des.* **13,** 3047–3058 (2007).
17. Höfer, T. *et al. Arch. Toxicol.* **78,** 549–564 (2004).
18. Hartung, T. *ALTEX* **24,** 143–147 (2007).
19. Hareng, L., Pellizzer, C., Bremer, S., Schwarz, M. & Hartung, T. *Reprod. Toxicol.* **20,** 441–452 (2005).
20. Kola, I. & Landis, J. *Nature Rev. Drug Discov.* **3,** 711–715 (2004).
21. Kuiper, H. A. *Plant J.* **27,** 503–528 (2001).
22. Service, R. F. *Science* **321,** 1036–1037 (2008).
23. Wood, M., Pichard, A., Gundert-Remy, U., de Rooij, C. & Tissot, S. M. *J. Hazard. Mater.* **133,** 8–15 (2006).
24. http://www.cochrane.org/
25. Hoffmann, S. & Hartung, T. *Hum. Exp. Toxicol.* **25,** 497–513 (2006).
26. Hansson, S. O. & Ruden, C. *Toxicology* **218,** 100–111 (2006).
27. Davila, J. C. *et al. Toxicol. Sci.* **79,** 214–223 (2004).
28. Stummann, T. C. & Bremer, S. *Curr. Stem Cell Res. Ther.* **3,** 117–130 (2008).
29. Hartung, T. & Leist, M. *ALTEX* **25,** 91–96 (2008).
30. Leist, M., Hartung, T. & Nicotera, P. *ALTEX* **25,** 103–114 (2008).
31. MacGregor, J. T. *Toxicol. Sci.* **75,** 236–248 (2003).
32. Pognan, F. *Pharmacogenomics* **5,** 879–893 (2004).
33. Lotti, M. & Nicotera, P. *Nature* **416,** 481 (2002).
34. Orrhenius, S. & Zhitvotovsky, B. *Chem. Res. Toxicol.* **19,** 729–733 (2006).
35. Committee on Toxicity Testing and Assessment of Environmental Agents, Board on Environmental Studies and Toxicology, Institute for Laboratory Animal Research, National Research Council of the National Academies. *Toxicity Testing in the 21st Century: a Vision and a Strategy* (The National Academies Press, 2007).
36. Collins, F. S., Gray, G. M. & Bucher, J. R. *Science* **319,** 906–907 (2008).
37. http://www.epa.gov/osa/spc/toxicitytesting/
38. http://www.epa.gov/ncct/toxcast/
39. Hartung, T. *ALTEX* **25,** 147–162 (2008).
40. Bottini, A. A. & Hartung, T. *ALTEX* **26,** 3–16 (2008).
41. Hartung, T. *Altern. Lab. Anim.* **36,** 635–639 (2008).
42. Hartung, T. *Toxicol. Sci.* **109,** 18–23 (2009).

REVIEWS

The role of black holes in galaxy formation and evolution

A. Cattaneo[1,2], S. M. Faber[3], J. Binney[4], A. Dekel[5], J. Kormendy[6], R. Mushotzky[7], A. Babul[8], P. N. Best[9], M. Brüggen[10], A. C. Fabian[11], C. S. Frenk[12], A. Khalatyan[13], H. Netzer[14], A. Mahdavi[15], J. Silk[4], M. Steinmetz[1] & L. Wisotzki[1]

Virtually all massive galaxies, including our own, host central black holes ranging in mass from millions to billions of solar masses. The growth of these black holes releases vast amounts of energy that powers quasars and other weaker active galactic nuclei. A tiny fraction of this energy, if absorbed by the host galaxy, could halt star formation by heating and ejecting ambient gas. A central question in galaxy evolution is the degree to which this process has caused the decline of star formation in large elliptical galaxies, which typically have little cold gas and few young stars, unlike spiral galaxies.

Galaxies come in two basic types: 'football-shaped' ellipticals and 'disk-shaped' spirals (Fig. 1). Spirals contain plenty of cold gas, which forms stars, whereas the gas in ellipticals is too hot to form stars. Thus, ellipticals lack the young blue stars that are usually seen in spirals, and are generally quite red. Spirals also have central bulges structurally resembling miniature ellipticals. Owing to this similarity, we use the term 'bulges' for bulges within spirals and for ellipticals indiscriminately.

Each bulge contains a central black hole, whose mass is proportional to the bulge stellar mass[1–5], $M_{BH} \approx 0.001 M_{bulge}$. Black holes and bulges also formed at about the same epoch in the lifetime of the Universe[6,7]. These observations imply that the formation of black holes and the formation of bulges are closely linked. Matter falling onto a black hole releases a huge amount of energy[8], of the order of 10% of the rest mass energy, $E = mc^2$, mainly in the form of photons but also in the form of radio-luminous jets of charged particles[9,10]. Even a tiny fraction (<1%) of the energy released within each bulge could heat and blow away its entire gas content, thus explaining the lack of star formation in bulges.

The theorist's goal is to understand these observations in a cosmological context. In the standard picture[11–13], most of the Universe is composed of dark matter, whose nature is unknown. Protons, electrons and neutrons, which compose gas and stars, make up the rest. They interact with dark matter purely through gravity, which determines the evolution of the Universe on large scales. The Universe emerged from the Big Bang with small inhomogeneities. These eventually grew into lumps, called haloes, by attracting surrounding matter gravitationally (Fig. 2). The competition between radiative cooling and gravitational heating determines the fate of gas in these haloes[14–16]. In low-mass haloes, cooling dominates. Galaxies grow through the accretion of gas that falls to the centre in cold flows[17,18], settles into disks[19] (but see refs 20, 21), and forms stars. However, when the halo mass grows above a critical value of about 10^{12} solar masses[18], heating dominates, and the gas no longer accretes onto galaxies. Halo mergers form large haloes that contain tens or even hundreds of galaxies, called groups or clusters, respectively. Galaxy mergers within haloes

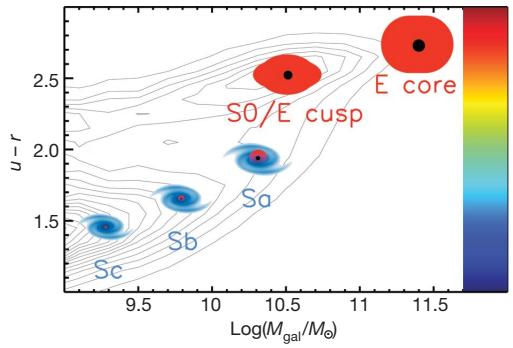


Figure 1 | The galaxy bimodality. The contours show the galaxy distribution on a stellar mass (M_{gal}) – colour diagram[92]. The difference between ultraviolet luminosity and red luminosity, quantified by the magnitude difference $u - r$, is a colour indicator; larger values of $u - r$ correspond to redder galaxies. The colour bar has been inserted to convey this notion visually and has no quantitative meaning. Galaxies are classified into two main types: spirals that mainly grew through gas accretion ('S', shown in blue) and ellipticals that mainly grew through mergers with other galaxies ('E', shown in red). 'S0' galaxies are an intermediate type, but we assimilate them to ellipticals. Spirals have central bulges, shown in red, that resemble miniature ellipticals. All ellipticals and bulges within spirals contain a central black hole, shown with a black dot. Moreover, ellipticals and bulges within spirals have the same black-hole mass to stellar mass ratio, of the order of 0.1%. This is why we call them 'bulges' indiscriminately. In contrast, there is no connection between masses of black holes and masses of disks (the galactic component shown in blue). Spirals and ellipticals are separated by a colour watershed at $u - r \approx 2$ and a mass watershed at $M_{gal} \approx M^* \approx 10^{10.5} M_\odot$ (ref. 92). M^* is of the order of $f_b M_{crit}$, where $M_{crit} \approx 10^{12} M_\odot$ is the critical halo mass for gas accretion and $f_b \approx 0.17$ is the cosmic baryon fraction. Spirals form a sequence where the bulge-to-disk ratio tends to grow with M_{gal} (Sc, Sb, Sa). Ellipticals have two subtypes[57,93]: giant ellipticals with smooth low-density central cores formed in mergers of galaxies that have long finished their gas ('E core')[94] and lower-mass ellipticals with steep central light cusps formed in mergers of galaxies that still have gas ('E cusp')[95]. Whereas core ellipticals formed all their stars over a short time span at high redshift[42], the formation of the lower-mass cuspy ellipticals from the 'quenching' and reddening of blue galaxies continues to low redshift[96].

[1]Astrophysikalisches Institut Potsdam, An der Sternwarte 16, 14482 Potsdam, Germany. [2]Observatoire de Lyon, Université de Lyon 1, 9 avenue Charles André, 69561 Saint Genis Laval cedex, France. [3]University of California Observatories/Lick Observatory, University of California, Santa Cruz, California 95064, USA. [4]Department of Physics, University of Oxford, Keble Road, Oxford OX1 3RH, UK. [5]Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel. [6]Department of Astronomy, University of Texas, Austin, Texas 78712, USA. [7]Goddard Space Flight Center, NASA, Greenbelt, Maryland 20771, USA. [8]Department of Physics and Astronomy, University of Victoria, Elliot Building, 3800 Finnerty Road, Victoria, British Columbia V8P 1A1, Canada. [9]Institute for Astronomy, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK. [10]Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany. [11]Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK. [12]Institute for Computational Cosmology, University of Durham, South Road, Durham DH1 3LE, UK. [13]Observatoire Astronomique Marseille-Provence, 38 rue Frédéric Joliot-Curie, 13388 Marseille cedex 13, France. [14]Wise Observatory, University of Tel Aviv, 69978 Tel Aviv, Israel. [15]Department of Physics and Astronomy, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132, USA.

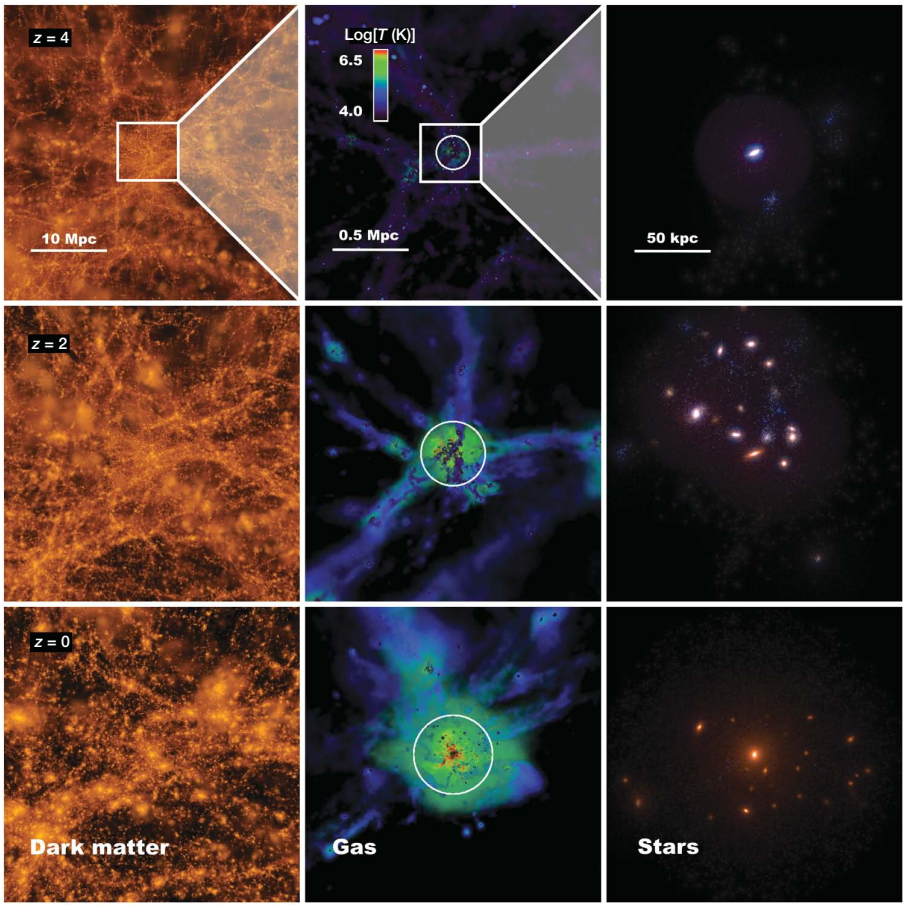


Figure 2 | A computer simulation of the formation of an elliptical galaxy. The nine panels illustrate the formation of an elliptical galaxy[97] by showing how the dark matter (left column), the gas (centre column) and the stars (right column) are distributed at three epochs in the expansion of the Universe: when the Universe was 1/5 of its current size (redshift $z = 4$), when the Universe was 1/3 of its current size ($z = 2$), and today ($z = 0$). The gravity of the dark matter dominates the evolution on large scales (left column). As time passes, the Universe becomes lumpier because the dark matter clumps via gravity into haloes (bright orange spots in the left panels). The centre column zooms into the region around and inside a halo to show what happens to the gas. The halo radius is shown as a white circle, and the gas is colour-coded according to its temperature: blue is cold, green (and red) is hot. Initially the halo is small, and the gas streams into the halo down to its centre in cold flows. When the halo reaches the critical mass $M_{crit} \approx 10^{12} M_{\odot}$ ($z = 2$), the gas begins to form a hot atmosphere (green); eventually, all the gas within the halo is hot ($z = 0$). The right column zooms in even further to show the visible galaxy formed by the gas fallen to the centre. The galaxy is initially a blue spiral ($z = 4$). It starts to become red when the halo gas starts to be hot ($z = 2$). By then, its halo has merged with neighbouring haloes to form a galaxy group. Mergers with companions eventually transform the galaxy into an elliptical ($z = 0$).

transform disks into bulges[22] and are the only opportunity for galaxies to grow after they have ceased to accrete gas.

In mergers of galaxies that are still accreting gas, the gas falls to the centre, triggers starbursts, and is often observed to feed the rapid growth of black holes[23,24], although, observationally, the connection between active galactic nuclei (AGN) and mergers remains controversial[25–27]. Black holes respond to this fuelling by feeding energy back to the surrounding gas. This energy produces winds[28–33], which may accelerate the star formation rate by compressing the gas[34]. In the most dramatic scenario, all the gas is blown away, so black hole growth and star formation suddenly terminate[35–39]. Computer simulations suggest that this 'quenching' is necessary to explain why ellipticals are red[40,41]. The chemical composition of giant ellipticals implies that the star formation epoch was brief. This is another reason for considering a mechanism that could 'quench' star formation abruptly[42].

In galaxies that have ceased to accrete gas, the main epoch of star formation and black hole accretion is over, but even weak AGN can open large cavities in the hot gas through the mechanical action of their jets[43–45]. In this 'maintenance mode', black holes couple to the hot gas and adjust their residual accretion rates to provide the energy that is needed to maintain it at constant temperature[46–48], either through a continuous series of minor events[44,45,49], or through episodic quasar activity[50,51]. The strongest evidence for this loop cycle is in galaxy clusters. Its action reduces considerably the rate at which gas cools and accretes onto the central galaxies[52,53].

These discoveries have lead to a richer picture of galaxy formation, in which black holes play a major part. Modelling these effects is essential to understand the masses, colours[54–56] and structural properties[57] of ellipticals within a cosmological theory of galaxy formation.

Black holes in the formation of red ellipticals

Black hole masses are tightly related to the stellar mass[3] and the stellar velocity dispersion[4,5] within the host bulges. These correlations imply a causal link between the formation of black holes and the formation of bulges, but they can be interpreted in two ways.

In the first interpretation, black hole accretion and star formation occur together because they both feed from the same gas, brought to the centre by gas-rich mergers and disk instabilities. Black hole accretion terminates when star formation has used up all the gas. The correlation between black hole growth and starbursts observed in ultraluminous infrared galaxies[24], quasars[58] and nearby Seyfert galaxies[59] supports this scenario, explaining why the black hole mass is proportional to the bulge mass.

In the second interpretation, star formation terminates when the black hole blows all the gas outside its host galaxy[35–38]. Feedback

requires a minimum power and thus a minimum mass because, for a given black hole mass, there is a maximum AGN luminosity, called the Eddington limit, above which the radiation-pressure force outwards exceeds the gravitational force inwards, suppressing the gas flow onto the black hole. The velocity dispersion is the bulge property that is most closely linked to the black hole because it determines the depth of the potential well from which the gas has to be expelled, and thus the minimum black hole mass for feedback.

The efficiency with which AGN can transfer energy to the surrounding gas determines which picture is closer to reality. Photons and jets from AGN can drive winds in two ways. They can heat the gas and cause it to expand (thermal 'energy-driven' winds) or they can push it out (pressure 'momentum-driven' winds). Photons heat the gas by photoionizing metals such as iron, which retain their inner electrons even at high temperature, and by Compton scattering. Radiation pressure on ionized gas is only important in the immediate proximity of black holes, where it determines the Eddington limit, but radiation pressure on dust is important even on galactic scales because dust has a high absorption cross-section. Absorption by resonance lines is another mechanism that transfers momentum to the gas, and may explain the high-speed winds in broad absorption line quasars[33]. Jets can produce 'energy-driven' winds via shock heating and 'momentum-driven' winds via ram pressure.

All these processes contain large inefficiencies, which are difficult to quantify: metals that retain some electrons even at high temperature are a small fraction of the atoms in the gas; the photoionization cross-section is large only in a narrow frequency band; Compton scattering transfers only a small fraction of the photon energy to an electron; and jets tend to escape from their host galaxies and to deposit most of their energy outside. The transfer of momentum from the photons to the gas by way of radiation pressure is the only process that can be almost 100% efficient, but dust must cover a large solid angle. The low momentum of photons with respect to their energy also limits the mass that can be ejected through this mechanism, as photons are massless particles. If the momentum in the photons radiated by an AGN was transferred to the gas with 100% efficiency, this momentum could eject a gas mass equal to at most ~10% of the bulge stellar mass, which is the mass of the gas that is typically left over at the end of a gas-rich merger. Thermal-wind[35,39] and radiation-pressure-driven-wind[36–38] models have been used to compute the relation between black hole mass and bulge velocity dispersion. However, this is not a discriminating test because both cases can match the data.

Optical/ultraviolet[28,29] and X-ray[30,31] spectroscopy confirm that quasars can accelerate winds to speeds of thousands to tens of thousands of kilometres per second. Their variability on short timescales suggests that these are nuclear rather than galactic winds, but some kiloparsec-scale winds are observed[29]. Integral-field spectroscopy has also detected bipolar winds with speeds of thousands of kilometres per second aligned with the jets of high-redshift radio galaxies[32].

The mere existence of AGN winds is no proof that they 'quench' star formation, but observations of post-starburst galaxies find that two-thirds of them contain winds with speeds of 500–2,000 km s^{-1} (ref. 60). These speeds are higher than the wind speeds usually found in starbursts. They, therefore, suggest a quasar origin and a probable role of quasars in the 'quenching' of post-starburst galaxies.

In the Sloan Digital Sky Survey (SDSS), which probes nearby galaxies, star-forming, 'active' and 'passive' ellipticals delineate a sequence from blue to red on the galaxy colour–mass diagram[61]. The interpretation is that when the growth of the black hole is activated, the star formation rate declines. However, in the overall population of SDSS galaxies, the star formation decline appears to be gradual[62] and is not linked to any dramatic event. The situation may be different at high redshift, where starbursts and quasars were more common and more powerful, and where star formation in the progenitors of giant ellipticals lasted for less than a gigayear (ref. 42).

Black holes in galaxy clusters

The gas in massive galaxies, groups and clusters is hot and radiates copiously in X-rays. The problem of explaining why this gas does not quickly cool off is known as the 'cooling flow' problem. This problem has been heavily investigated in galaxy clusters, where the observational constraints are particularly strong[52,53].

X-ray groups and clusters fall into two categories: systems in which the X-ray surface brightness increases steeply towards the centre—that is, 90% of the X-ray-selected groups and clusters with halo mass (M_{halo}) $\leq 10^{14}$ solar masses, and 50% of the clusters with $M_{halo} \geq 10^{14}$ solar masses[63]—and systems with shallower surface brightness profiles. The gas luminosity per unit volume is equal to $n^2\Lambda$, where n is the gas density, and Λ is the cooling function; Λ depends on the gas temperature T and on the gas chemical composition. X-ray spectroscopy shows that T is always of the order of the virial temperature T_{vir} at which the gas is in equilibrium with gravity. Therefore, the difference between the two types of cluster must be in their density profiles.

In the first type of cluster, the gas has a high central density and radiates its thermal energy on a timescale $t_{cool} \approx (3/2)kT/n^2\Lambda$, where k is the Boltzmann constant, that is usually less than a gigayear over much of the cluster core. These clusters are called 'cool-core' clusters because T decreases towards the centre. However, cool-core clusters are not 'cold core' clusters: the temperature drops towards the centre by only a factor of three. From the absence or weakness of the soft X-ray line Fe XVII, one infers that the amount of gas that cools radiatively below this temperature is ten times less than expected from how much heat is lost to X-rays[52,53]. As the gas radiates but does not cool, there must be a compensating energy-injection mechanism.

Further evidence comes from the relation between X-ray luminosity (L_X) and the gas temperature measured from X-ray data. At $T \geq 3$ keV, where bremsstrahlung is the main radiation mechanism, the X-ray luminosity is $L_X \propto n^2 {T_{vir}}^{1/2} {r_{halo}}^3$, and $T_{vir} \propto M_{halo}/r_{halo}$, where r_{halo} is halo radius. If n scaled with the halo density, which is proportional to the mean density of the Universe, then all clusters should have the same n and this equation predicts $L_X \propto {T_{vir}}^2$. Spectroscopy confirms that $T \approx T_{vir}$, but the data find a different relation $L_X \propto {T_{vir}}^3$, which becomes even steeper at $T_{vir} \leq 3$ keV (ref. 64) because n decreases at low masses, although with considerable scatter[65–67]. Having lower density for a given temperature implies having higher entropy, measured by $K = kT/n^{2/3}$, and the only way to increase the entropy is through heating.

Clusters have typical entropy excesses of $\Delta K \approx 100$ keV cm^2 at $0.1r_{halo}$ (ref. 68; Fig. 3). These excesses weigh more heavily on smaller clusters, which have lower absolute entropies, but higher entropies relative to theoretical expectations. This problem is common to both cool-core and non-cool-core clusters, and affects a large fraction of the intracluster medium. The quasar winds invoked to quench star formation in the progenitors of giant ellipticals could solve this entropy problem by preheating the intergalactic gas destined to become the intracluster medium[69–71], but they cannot solve the cooling-flow problem in the central regions of cool-core clusters. In these systems, which have $K < 100$ keV cm^2, the cooling time is so short (<0.1 Gyr; Fig. 3) that heating is needed at least every 0.1 Gyr today to maintain these systems in their current state. This need for regular heating clashes with the scarcity of quasars in the low-redshift Universe.

However, weaker AGN—that is, 'edge-darkened' radio galaxies—show up in 70% of the central dominant (cD) galaxies of cool-core clusters[72,73]. Their activity pattern differs from the erratic behaviour of quasars and is closer to a constant string of minor outbursts. In many low-accretion-rate AGN, the gas surrounding the black hole is not dense enough to radiate efficiently, and nearly all the released energy may instead be channelled into jets[74]. The very poor optical luminosity of these objects is partly the reason why their importance had long been underestimated.

The importance of radio galaxies began to emerge after cavities were discovered in the X-ray gas of the Perseus cluster[43,45] (Fig. 4) and other clusters with substantially weaker AGN—for example, the Virgo

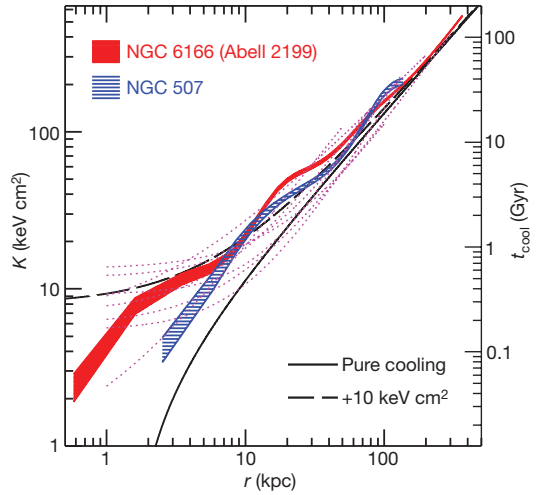


Figure 3 | The entropy of the intracluster medium in spherical shells of radius *r*. We define entropy as $K = kTn^{1/(\gamma-1)}$, where γ is the gas adiabatic index. For an ideal gas, this definition is related to the standard one, $\Delta S = \Delta Q/T$, by the transformation $s = \ln K^{1/(\gamma-1)} + \text{const.}$, where s is the entropy S per unit mass (ΔS is the entropy variation that corresponds to a heat injection ΔQ). Observed entropy profiles of cool-core clusters[98] (red dotted lines) differ substantially from theoretical predictions for a pure cooling flow model (black solid line) but become broadly consistent with theoretical predictions if an entropy pedestal of 10 keV cm^2 is added to the latter (dashed black line). The discrepancy with predictions of pure cooling flow models is even larger in non-cool-core clusters, which have central entropies up to 700 keV cm^2 (for example, 3C 129). A new study by three of us (A.M., A.B. and A.C., unpublished results) resolves entropy profiles of NGC 507 and NGC 6166, the cD galaxy of A 2199, at small radii and finds that the pedestal is actually a shelf. The entropy decreases again at small radii. Both galaxies have half-light radii of ~10 kpc. We also show the radiative cooling times t_{cool} that approximately correspond to the entropies on the *y*-axis of the diagram. The dependence of t_{cool} on K, $t_{cool} = 3/2K^{3/2}/(\Lambda\sqrt{kT})$ for $\gamma = 5/3$, is stronger than that on T because of the exponent 3/2 and because the range of entropies within a cluster and among clusters is much larger than the corresponding range of temperatures.

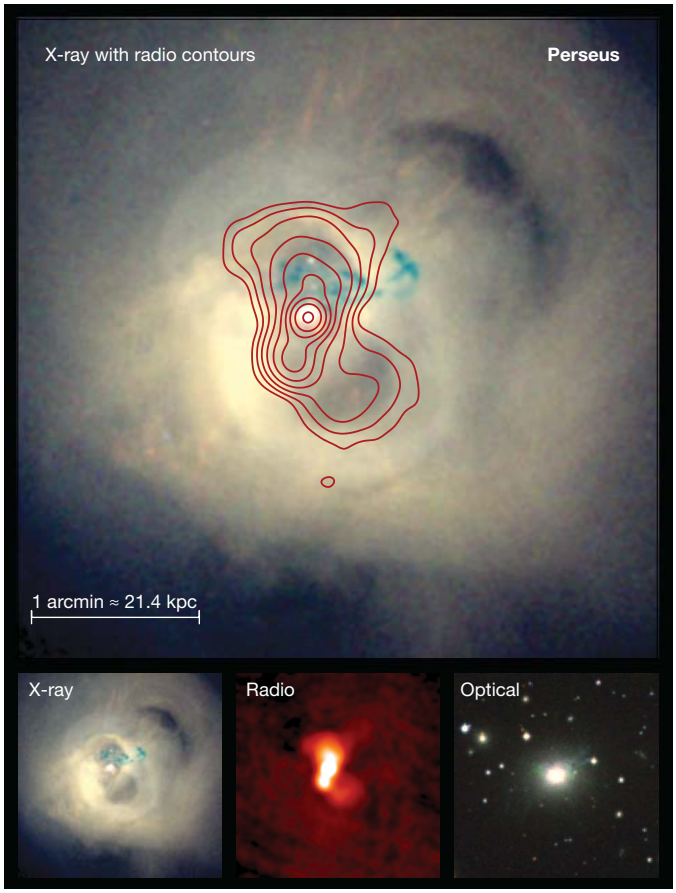


Figure 4 | Optical, radio and X-ray images of the Perseus cluster. The optical image (bottom right panel) shows the Perseus galaxy cluster and its cD galaxy, the radio galaxy Perseus A. The radio images (bottom middle panel, and contours on main panel) show the lobes of relativistic synchrotron emitting plasma inflated by the jets that come out of the central nucleus. The X-ray images (main panel and bottom left panel) show the ambient hot gas (intracluster medium), colour-scaled such that higher-energy X-rays (hotter gas) are bluer and lower-energy X-rays (cooler gas) are redder (the X-ray and radio data are from ref. 45). This is one of the most dramatic images of the cavities and ripples created by AGN in the surrounding gas. The regions of higher radio luminosity are dimmer in X-rays because the expanding lobes have displaced the ambient gas. The rims of the radio lobes are X-ray bright and are cooler than the surrounding gas. This is more obvious if one looks to the left of the spot where the jet directed downward terminates. A third X-ray cavity has the shape of an arch and is visible at the top right corner. This is called a 'ghost' cavity because it is invisible in radio, and it was generated during a previous burst of activity. The pink brush strokes running perpendicular to the arch are cold gas flowing around the cavity. The blue structure to the north of the active nucleus is due to absorption in an infalling system in front of Perseus A. This feature appears because higher-energy X-rays are more penetrating. Images from ref. 45.

cluster[44]. The cavities are regions where the jet radio-synchrotron-emitting plasma has displaced the ambient X-ray-emitting plasma. Such cavities are present in ≥70% of cool-core clusters[75]. They are usually regions of enhanced synchrotron emission, although some lack high-frequency radio emission ('ghost' cavities), presumably because they are old and depleted of electrons with energies $>10^5 m_e c^2$, where m_e is the electron mass.

From the volume of the cavities (V_{cav}) and the pressure of the intracluster medium (p_{ICM}) it is possible to estimate the work that the jets had to do to create them. This work equals $p_{ICM}V_{cav}$ for 'quasi-static' (that is, highly subsonic) inflation. However, as cavities are Rayleigh-Taylor unstable, their formation timescale cannot be much longer than the sound crossing time or they would break apart before they are formed. It cannot be much shorter than the sound crossing time either. If it were, cavities would be surrounded by strong shocks, observed only in very few objects. As cavities are not inflated quasi-statically, the work that the expanding radio lobes do on the ambient gas must exceed $p_{ICM}V_{cav}$. Part of this extra work excites shocks, waves and other disturbances, all of which could heat the intracluster medium. Analytic calculations[76] and numerical simulations[77] indicate that this work could be up to $10p_{ICM}V_{cav}$.

In addition to the work that radio lobes do on the ambient gas, there is also the energy of the relativistic particles inside the radio lobes themselves ('cosmic rays'). The internal energy of the radio-emitting plasma is $1/(\gamma-1)p_{cav}V_{cav}$, where cavity pressure $p_{cav} \geq p_{ICM}$ and γ is the plasma adiabatic index ($\gamma = 4/3$ is the value that is normally assumed for a relativistic plasma). This energy could heat the intracluster medium, too, if the synchrotron emitting plasma and the X-ray emitting gas eventually mixed, although observations show that cavities tend to survive in the intracluster medium for a very long time (see, for example, ref. 75).

The minimum energy needed to produce the observed cavities (E_{cav}) is obtained by adding the work done on the ambient gas for quasi-static inflation and the cosmic ray energy. This sum gives $E_{cav} \geq 4p_{ICM}V_{cav}$. Cavity observations find that this energy is equal to the energy radiated in X-rays in a sound crossing time, to within a factor of four (ref. 46; Fig. 5a). Thus, the energy that jets put into cavities is about equal to the energy needed to offset cooling. This near-equality, which extends over four orders of magnitude, suggests a self-regulation mechanism. This is possible, because a black hole acts as a thermostat that senses the entropy of the gas at the boundary of its gravitational sphere of influence, determined by the Bondi

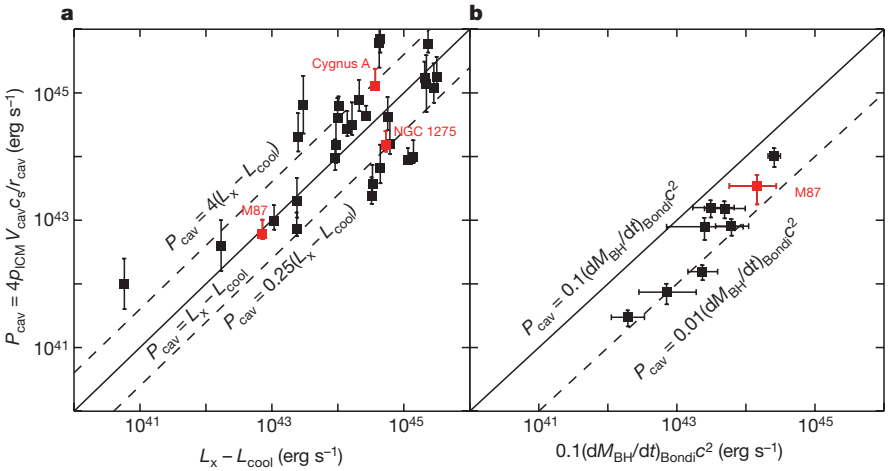


Figure 5 | Cooling, heating, and black hole accretion rates. a, P_{cav} measures the mechanical energy E_{cav} used to create a cavity divided by the sound crossing time r_{cav}/c_s (ref. 46). L_X is the total X-ray luminosity of the intracluster medium inside the cooling radius (defined[46] as the radius within which $t_{cool} < 7.7$ Gyr). L_{cool} is the X-ray luminosity of the gas that is actually cooling, computed by modelling X-ray and ultraviolet spectra. In fact, L_{cool} is negligible compared to L_X because the gas-cooling rate is low. If $E_{cav} = 4p_{ICM}V_{cav}$, then $P_{cav} = L_X$ to within a factor of four. This means that the jet energy is about equal to the energy that is needed to offset cooling. **b**, P_{cav} ranges[79] from 10% to 100% of the accretion power determined from the Bondi spherical accretion model[78]. This shows that the Bondi model is a reasonable description of the accretion of hot gas by a supermassive black hole. The different ordinates of M87 in **a** and **b** show that the published error bars underestimate the real uncertainty on P_{cav}. Error bars, 1σ.

radius[78] $r_{Bondi} \approx GM_{BH}/c_s^2 \approx 10$–$100$ pc, where the gas infall speed equals the sound speed c_s (here G is the gravitational constant). For spherical accretion[78], M_{BH} and the entropy at r_{Bondi} entirely determine the black hole accretion rate. If the power injected into the cavities is proportional to the black hole accretion rate computed from a spherical model—an assumption directly verified by observational data[79] (Fig. 5b)—then, the more the gas cools down and the central entropy decreases, the more the jet power and the heating rate go up.

The importance of jet heating is also shown by the entropy of the intracluster medium, which is higher in clusters with extended radio sources than in clusters with point-like radio sources, where jets have not yet propagated and heated the gas[80]. Properties of clusters with point-like sources are in fact consistent with the non-heating prediction, $L_X \propto T_{vir}^2$.

The problem is how the energy injected into clusters is converted into heat. Strong shocks seem the most natural mechanism, but in fact the X-ray-bright rims of the radio lobes in the Perseus cluster[45] (Fig. 4) and other radio galaxies are cooler than their surroundings. Moreover, if the heated region were substantially smaller than the cluster core, convection would set in. Instead, the entropy profiles of clusters are shallow but not flat (Fig. 3). This indicates that the central gas is stable against convection, even though the gas metallicity profiles are broader than the starlight of the cD galaxies[53]. Different stages in the life of a radio source could explain the scarcity of radio sources with strong shocks and the need for distributed heating[81]. In the transient active phase, jets inflate cavities and shocks are the main heating mechanism. If the jets can cross the cluster core before this phase ends, then shock heating raises the entropy by a nearly uniform amount throughout the core, in agreement with the observed $\sim$10 keV cm^2 entropy pedestal (Fig. 3). After the jets switch off, the radio lobes keep doing mechanical work on the intracluster medium by rising buoyantly. The thermalization of the hydrodynamic motions generated during the active phase and of the waves generated by the rising bubbles generates heat at various radii, even when the black hole is not actually accreting[44,45,49,76].

Heating is not the only mechanism by which AGN can prevent cold gas from accumulating at the centres of clusters. Jets and rising bubbles also lift low-entropy gas from the central region and transport it outside. Eventually this gas falls down again, but new bubbles are created and lift it up again[82]. This would explain the filaments of cold molecular gas detected around the cavities of the Perseus cluster[83].

Black holes in galaxy evolution

Giant ellipticals have the same cooling flow problem as galaxy clusters, with even stronger limits on the amount of gas that can cool and form stars. The stellar populations, chemical abundances, and structural properties (that is, the absence of dense central light cusps[57]) of giant ellipticals indicate that little gas has fallen to the centre and made stars since these galaxies were formed.

The cooling flow problem is more severe within galaxies than within clusters because, even neglecting the hot gas in the halo, the final stages of the lives of massive stars return $\sim$30–40% of the total stellar mass to the interstellar medium over the lifetime of the Universe[50,51,84]. Even a small fraction of the gas from dying massive stars would, if accreted, result in black holes much larger than the observational mass estimates.

The problem with applying to galaxies the same explanation as applied to clusters is that jets are usually collimated on galactic scales[85] (see, for example, M87). Therefore, they drill through the nearby gas and dump most of the energy outside the galaxies in which they are produced: the entropy shelf surrounding NGC 6166 (Fig. 3) indicates that heating has only been important at radii $r \geq 2$ kpc. Even in the Perseus cluster, the jets seem to not be inhibiting star formation in the central galaxy, which belongs to the 25% of cD galaxies that are blue[86]. The situation is even worse in galaxies that are not at the centres of clusters, because confining the jets is even more difficult for the less pressurized atmospheres of these galaxies. Without a confining working surface, jets dissipate their energy uselessly in intergalactic space[57] (for example, Cygnus A). However, there are counter-examples where the jets have caused turmoil in the hot gas on galactic scales—for example, Centaurus A, M84 and NGC 3801. Moreover, a jet may escape from its host galaxy and still transfer some of its energy to the interstellar medium. For example, the knots in the jet of M87 could be interpreted as evidence for interaction with the interstellar medium. Despite this problem, it is intriguing that the fraction of ellipticals hosting a radio source scales with M_{BH} in the same way as does the estimated gas cooling rate[87], and that the time-averaged jet power matches the gas X-ray luminosity over two orders of magnitude in galaxy mass[88]. If jets fail to couple to the ambient gas and to keep it hot, cooling will eventually activate an optical AGN, which could heat the gas radiatively[51].

Gravitational heating, due to the mechanical work done by infalling clumps when they fall deep into the galaxies[89–91], also contributes to heating the gas, as do type I supernovae in small ellipticals[50,51,84].

However, these energy sources are unresponsive to changes in the radiative loss rate. Either they will heat the gas at a rate lower than the cooling rate, in which case the gas will eventually cool, or they will heat the gas at a rate higher than the cooling rate, in which case they will drive an outflow. The gas may be in an outflow in lower-mass ellipticals (the 'cuspy' ellipticals of Fig. 1), where discrete sources dominate the X-ray emission and where, for this reason, we cannot generally detect any X-ray-emitting gas, but in giant ellipticals X-ray observations show that the gas is in hydrostatic equilibrium[57]. Gravitational heating and type I supernovae could, nevertheless, alleviate the burden on the AGN, which would only provide the difference between the heating rate needed to keep the gas in equilibrium and the heating rate provided by these other sources.

Remaining issues

The strongest evidence for black hole feedback is in galaxy clusters, but we still lack a sufficient understanding of the processes that transfer energy from AGN to the surrounding gas and thermalize the hydrodynamic disturbances excited by expanding jets and rising bubbles. Standard viscosity, turbulent viscosity, the stretching and tearing of magnetic field lines, and cosmic rays could all contribute to heat and/or lift the intracluster medium. The statement that black holes self-regulate to the accretion rate that is required to offset cooling is a valid first approximation, but some gas does cool and flow onto the central galaxies of clusters[83], although at a very low rate compared to predictions for pure cooling flow models. In 25% of all clusters, this gas reactivates star formation, leading to blue-core cD galaxies[86]. A major challenge for theoretical models and computer simulations is to understand in quantitative detail why real clusters depart from an 'ideal' feedback loop that is 100% efficient in suppressing cooling and star formation.

The interaction of radio galaxies with their own interstellar medium is much less clear than the interaction with the intracluster medium. In the case of radiative feedback, the basic physics of the interaction with the interstellar medium are much better understood. The main open problem is rather whether radiative feedback can deliver the energy required for the 'maintenance' of individual ellipticals without exceeding the observational limits on the fraction containing an AGN.

The greatest uncertainty is the role of quasar winds in quenching star formation. This is because the masses of the winds detected spectroscopically are uncertain by more than one order of magnitude. Improving the current estimates for the masses, length scales, and temperature structure of the winds at all redshifts is the critical observational challenge. We also need to understand better the properties of galaxies in transition from the blue to the red population.

Finally, we note that it is computer simulations that indicate the need for quasar quenching, but these simulations are based on uncertain models for star formation and the physics of the interstellar medium. Progress in our understanding of these processes and higher resolution simulations will be necessary before we can conclude that quasar feedback is in fact needed, particularly in lower-mass ellipticals where the decline of the star formation rate occurs on a longer timescale.

1. Kormendy, J. in *The Nearest Active Galaxies* (eds Beckman, J., Colina, L. & Netzer, H.) 197–218 (Consejo Superior de Investigaciones Científicas, Madrid, 1993).
2. Magorrian, J. *et al.* The demography of massive dark objects in galaxy centers. *Astron. J.* **15,** 2285–2305 (1998).
3. Marconi, A. & Hunt, L. The relation between black hole mass, bulge mass, and near-infrared luminosity. *Astrophys. J.* **589,** 21–24 (2003).
4. Ferrarese, L. & Merritt, D. A fundamental relation between supermassive black holes and their host galaxies. *Astrophys. J.* **539,** 9–12 (2000).
5. Gebhardt, K. *et al.* A relationship between nuclear black hole mass and galaxy velocity dispersion. *Astrophys. J.* **539,** 13–16 (2000).
6. Cattaneo, A. & Bernardi, M. The quasar epoch and the stellar ages of early-type galaxies. *Mon. Not. R. Astron. Soc.* **344,** 45–52 (2003).
7. Hopkins, P. F. Determining the properties and evolution of red galaxies from the quasar luminosity function. *Astrophys. J.* **163** (Suppl.), 50–79 (2006).
8. Bardeen, J. M. Kerr metric black holes. *Nature* **226,** 64–65 (1970).
9. Krolik, J. H. *Active Galactic Nuclei: From the Central Black Hole to the Galactic Environment* (Princeton Univ. Press, 1999).
10. Cattaneo, A. & Best, P. N. On the jet contribution to the AGN cosmic energy budget. *Mon. Not. R. Astron. Soc.* **395,** 518–523 (2009).
11. White, S. D. M. & Rees, M. J. Core condensation in heavy halos — a two-stage theory for galaxy formation and clustering. *Mon. Not. R. Astron. Soc.* **183,** 341–358 (1978).
12. Blumenthal, G. R., Faber, S. M., Primack, J. R. & Rees, M. J. Formation of galaxies and large-scale structure with cold dark matter. *Nature* **311,** 517–525 (1984).
13. White, S. D. & Frenk, C. S. Galaxy formation through hierarchical clustering. *Astrophys. J.* **379,** 52–79 (1991).
14. Rees, M. J. & Ostriker, J. P. Cooling, dynamics and fragmentation of massive gas clouds — clues to the masses and radii of galaxies and clusters. *Mon. Not. R. Astron. Soc.* **179,** 541–559 (1977).
15. Silk, J. On the fragmentation of cosmic gas clouds. I — The formation of galaxies and the first generation of stars. *Astrophys. J.* **211,** 638–648 (1977).
16. Binney, J. The physics of dissipational galaxy formation. *Astrophys. J.* **215,** 483–491 (1977).
17. Keres, D. *et al.* How do galaxies get their gas? *Mon. Not. R. Astron. Soc.* **363,** 2–28 (2005).
18. Dekel, A. & Birnboim, Y. Galaxy bimodality due to cold flows and shock heating. *Mon. Not. R. Astron. Soc.* **368,** 39–55 (2006).
19. Fall, S. M. & Efstathiou, G. Formation and rotation of disc galaxies with haloes. *Mon. Not. R. Astron. Soc.* **193,** 189–206 (1980).
20. Steinmetz, M. Numerical simulations of galaxy formation. *Astrophys. Space Sci.* **269,** 513–532 (1999).
21. Dekel, A. *et al.* Cold streams in early massive hot haloes as the main mode of galaxy formation. *Nature* **457,** 451–454 (2009).
22. Toomre, A. & Toomre, J. Galactic bridges and tails. *Astrophys. J.* **178,** 623–666 (1972).
23. Sanders, D. B. *et al.* Ultraluminous infrared galaxies and the origin of quasars. *Astrophys. J.* **325,** 74–91 (1988).
24. Nardini, E. *et al.* Spectral decomposition of starbursts and AGNs in 5–8 micron Spitzer IRS spectra of local ULIRGs. *Mon. Not. R. Astron. Soc.* **385,** 130L–134L (2008).
25. Hutchings, J. B. & Campbell, B. Are QSOs activated by interactions between galaxies? *Nature* **303,** 584–588 (1983).
26. Dunlop, J. S. *et al.* Quasars, their host galaxies and their central black holes. *Mon. Not. R. Astron. Soc.* **340,** 1095–1135 (2003).
27. Bennert, N. *et al.* Evidence for merger remnants in early-type host galaxies of low-redshift QSOs. *Astrophys. J.* **677,** 846–867 (2008).
28. Arav, N. *et al.* HST STIS observations of PG0946+301: the highest quality spectrum of a BALQSO. *Astrophys. J.* **561,** 118–130 (2001).
29. de Kool, M. *et al.* Keck HIRES observations of the QSO FIRST J104459.6+365605: evidence for a large-scale outflow. *Astrophys. J.* **548,** 609–623 (2001).
30. Reeves, J. N., O'Brien, P. T. & Ward, M. J. A massive X-ray outflow from the quasar PDS456. *Astrophys. J.* **593,** 65–68 (2004).
31. Chartas, G., Brandt, W. N., Gallagher, S. C. & Proga, D. XMM-Newton and Chandra spectroscopy of the variable high-energy absorption of PG 1115+080: refined outflow constraints. *Astron. J.* **133,** 1849–1860 (2007).
32. Nesvadba, N. P. H. *et al.* Evidence for powerful AGN winds at high redshift: dynamics of galactic outflows in radio galaxies during the ''Quasar Era''. *Astron. Astrophys.* **491,** 407–424 (2008).
33. Proga, D. in *The Central Engine of Active Galactic Nuclei* (eds Ho, L. C. & Wang, J. M.) 267–276 (ASP Conf. Ser. Vol. 373, Astronomical Society of the Pacific, 2007).
34. Silk, J. Ultraluminous starbursts from supermassive black hole-induced outflows. *Mon. Not. R. Astron. Soc.* **364,** 1337–1342 (2005).
35. Silk, J. & Rees, M. J. Quasars and galaxy formation. *Astron. Astrophys.* **331,** 1L–4L (1998).
The first article to propose that the M_{BH}–σ is determined by feedback.
36. King, A. Black holes, galaxy formation, and the M_{BH}-σ relation. *Astrophys. J.* **596,** 27–29 (2003).
37. Murray, N., Quataert, E. & Thompson, T. A. On the maximum luminosity of galaxies and their central black holes: feedback from momentum-driven winds. *Astrophys. J.* **618,** 569–585 (2005).
38. Fabian, A. C., Celotti, A. & Erlund, M. C. Radiative pressure feedback by a quasar in a galactic bulge. *Mon. Not. R. Astron. Soc.* **373,** 16L–20L (2006).
Assuming that the growth of the black hole is limited by radiation pressure feedback yields the observed relation $M_{BH} \propto \sigma^4$.
39. Robertson, B. *et al.* The evolution of the M_{BH}-σ relation. *Astrophys. J.* **641,** 90–102 (2006).
40. Springel, V., Di Matteo, T. & Hernquist, L. Black holes in galaxy mergers: the formation of red elliptical galaxies. *Astrophys. J.* **620,** 79–82 (2005).
Without AGN feedback, the remnants of gas-rich mergers remain blue for several gigayears.
41. Hopkins, P. F. *et al.* A unified, merger-driven model of the origin of starbursts, quasars, the cosmic X-ray background, supermassive black holes, and galaxy spheroids. *Astrophys. J.* **163** (Suppl.), 1–49 (2006).
42. Thomas, D., Maraston, C., Bender, R. & Mendes de Oliveira, C. The epochs of early-type galaxy formation as a function of environment. *Astrophys. J.* **621,** 673–694 (2005).
43. Böhringer, H. *et al.* A ROSAT HRI study of the interaction of the X-ray-emitting gas and radio lobes of NGC 1275. *Mon. Not. R. Astron. Soc.* **264,** 25L–28L (1993).

44. Forman, W. *et al.* Reflections of active galactic nucleus outbursts in the gaseous atmosphere of M87. *Astrophys. J.* **635,** 894–906 (2005).
45. Fabian, A. C. *et al.* A very deep Chandra observation of the Perseus cluster: shocks, ripples and conduction. *Mon. Not. R. Astron. Soc.* **366,** 417–428 (2006).
Refs 43–45 show direct images of how radio galaxies can affect the intracluster gas.
46. Rafferty, D. A., McNamara, B. R., Nulsen, P. E. J. & Wise, M. W. The feedback-regulated growth of black holes and bulges through gas accretion and starbursts in cluster central dominant galaxies. *Astrophys. J.* **652,** 216–231 (2006).
The energy that black holes inject into the lobes of radio galaxies is approximately equal to the energy that is needed to offset cooling.
47. Churazov, E., Sunyaev, R., Forman, W. & Boehringer, H. Cooling flows as a calorimeter of active galactic nucleus mechanical power. *Mon. Not. R. Astron. Soc.* **332,** 729–734 (2002).
Black holes sense the entropy of the surrounding gas and adjust their accretion rate to offset cooling.
48. Cattaneo, A. & Teyssier, R. AGN self-regulation in cooling flow clusters. *Mon. Not. R. Astron. Soc.* **376,** 1547–1556 (2007).
49. Ruszkowski, M., Brueggen, M. & Begelman, M. Cluster heating by viscous dissipation of sound waves. *Astrophys. J.* **611,** 158–163 (2004).
50. Binney, J. & Tabor, G. Evolving cooling flows. *Mon. Not. R. Astron. Soc.* **276,** 663–678 (1995).
First introduction of black hole feedback as a possible solution to the cooling flow problem.
51. Ciotti, L. & Ostriker, J. P. Radiative feedback from massive black holes in elliptical galaxies: AGN flaring and central starburst fueled by recycled gas. *Astrophys. J.* **665,** 1038–1056 (2007).
Standard reference for radiative feedback and its role in the maintenance of elliptical galaxies.
52. Peterson, J. R. & Fabian, A. C. X-ray spectroscopy of cooling clusters. *Phys. Rep.* **427,** 1–39 (2006).
53. McNamara, B. R. & Nulsen, P. E. J. Heating hot atmospheres with active galactic nuclei. *Annu. Rev. Astron. Astrophys.* **45,** 117–175 (2007).
54. Croton, D. *et al.* The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. *Mon. Not. R. Astron. Soc.* **365,** 11–28 (2006).
55. Bower, R. *et al.* Breaking the hierarchy of galaxy formation. *Mon. Not. R. Astron. Soc.* **370,** 645–655 (2006).
56. Cattaneo, A. *et al.* Modelling the galaxy bimodality: shutdown above a critical halo mass. *Mon. Not. R. Astron. Soc.* **370,** 1651–1665 (2006).
Refs 54–56 show that the black hole–galaxy co-evolution contains an initial rapid growth phase followed by a 'maintenance' phase, in which black holes couple to the hot gas to suppress star formation.
57. Kormendy, J., Fisher, D. B., Cornell, M. E. & Bender, R. Structure and formation of elliptical and spheroidal galaxies. *Astrophys. J. Suppl. Ser.* (in the press); preprint at ⟨http://arxiv.org/abs/0810.1681⟩ (2008).
58. Lutz, D. *et al.* Star formation in the hosts of high-z QSOs: evidence from Spitzer PAH detections. *Astrophys. J.* **684,** 853–861 (2008).
59. Heckman, T. M. & Kauffmann, G. The host galaxies of AGN in the Sloan Digital Sky Survey. *N. Astron. Rev.* **50,** 677–684 (2006).
60. Tremonti, C. A., Moustakas, J. & Diamond-Stanic, A. M. The discovery of 1000 km s^{-1} outflows in massive poststarburst galaxies at z∼0.6. *Astrophys. J.* **663,** 77–80 (2007).
Most post-starburst galaxies contain winds of presumably AGN origin.
61. Schawinski, K. *et al.* Observational evidence for AGN feedback in early-type galaxies. *Mon. Not. R. Astron. Soc.* **382,** 1415–1431 (2007).
62. Quintero, A. *et al.* Selection and photometric properties of K+A galaxies. *Astrophys. J.* **602,** 190–199 (2004).
63. Chen, Y. *et al.* Statistics of X-ray observables for the cooling-core and non-cooling core galaxy clusters. *Astron. Astrophys.* **466,** 805–812 (2007).
64. Ponman, T. J., Bourner, P. D. J., Ebeling, H. & Boehringer, H. A. ROSAT survey of Hickson's compact galaxy groups. *Mon. Not. R. Astron. Soc.* **283,** 690–708 (1996).
65. Edge, A. C. & Stewart, G. C. EXOSAT observations of clusters of galaxies. I — The X-ray data. II — X-ray to optical correlations. *Mon. Not. R. Astron. Soc.* **252,** 414–441 (1991).
66. Evrard, A. E. & Henry, J. P. Expectations for X-ray cluster observations by the ROSAT satellite. *Astrophys. J.* **383,** 95–103 (1991).
67. Kaiser, N. Evolution of clusters of galaxies. *Astrophys. J.* **383,** 104–111 (1991).
68. Lloyd-Davies, E. J., Ponman, T. J. & Cannon, D. B. The entropy and energy of intergalactic gas in galaxy clusters. *Mon. Not. R. Astron. Soc.* **315,** 689–702 (2000).
69. Valageas, P. & Silk, J. The entropy history of the universe. *Astron. Astrophys.* **350,** 725–742 (1999).
70. Oh, S. P. & Benson, A. J. Entropy injection as a global feedback mechanism. *Mon. Not. R. Astron. Soc.* **342,** 664–672 (2003).
71. McCarthy, I. G., Babul, A., Bower, R. G. & Balogh, M. L. Towards a holistic view of the heating and cooling of the intracluster medium. *Mon. Not. R. Astron. Soc.* **386,** 1309–1331 (2008).
72. Burns, J. O. The radio properties of cD galaxies in Abell clusters. I — an X-ray selected sample. *Astron. J.* **99,** 14–30 (1990).
73. Best, P. N. *et al.* On the prevalence of radio-loud active galactic nuclei in brightest cluster galaxies: implications for AGN heating of cooling flows. *Mon. Not. R. Astron. Soc.* **379,** 894–908 (2007).
70% of cluster cD galaxies are radio galaxies: note X-ray cavities are present in a similar percentage of cool core clusters (ref. 76).
74. Blandford, R. D. & Begelman, M. C. On the fate of gas accreting at a low rate on to a black hole. *Mon. Not. R. Astron. Soc.* **303,** 1L–5L (1999).
75. Dunn, R. J. H. & Fabian, A. C. Investigating AGN heating in a sample of nearby clusters. *Mon. Not. R. Astron. Soc.* **373,** 959–971 (2006).
76. Nusser, A., Silk, J. & Babul, A. Suppressing cluster cooling flows by self-regulated heating from a spatially distributed population of active galactic nuclei. *Mon. Not. R. Astron. Soc.* **373,** 739–746 (2006).
77. Binney, J., Bibi, F. A. & Omma, H. Bubbles as tracers of heat input to cooling flows. *Mon. Not. R. Astron. Soc.* **377,** 142–146 (2007).
78. Bondi, H. On spherically symmetrical accretion. *Mon. Not. R. Astron. Soc.* **112,** 195–204 (1952).
79. Allen, S. W. *et al.* The relation between accretion rate and jet power in X-ray luminous elliptical galaxies. *Mon. Not. R. Astron. Soc.* **372,** 21–30 (2006).
80. Magliocchetti, M. & Brueggen, M. The interplay between radio galaxies and cluster environment. *Mon. Not. R. Astron. Soc.* **379,** 260–274 (2007).
The entropy of the intracluster medium is higher in clusters with extended radio sources.
81. Voit, G. M. & Donahue, M. An observationally motivated framework for AGN heating of cluster cores. *Astrophys. J.* **634,** 955–963 (2005).
82. Mathews, W. G. & Brighenti, F. Creation of X-ray cavities in clusters with cosmic rays. *Astrophys. J.* **660,** 1137–1145 (2006).
83. Salomé, P. *et al.* Cold molecular gas in the Perseus cluster core. *Astron. Astrophys.* **454,** 437–445 (2006).
84. Mathews, W. G. & Baker, J. C. Galactic winds. *Astrophys. J.* **170,** 241–260 (1971).
85. Helmboldt, J. F., Taylor, G. B., Walker, R. C. & Blandford, R. D. A statistical description of AGN jet evolution from the VLBA Imaging and Polarimetry Survey (VIPS). *Astrophys. J.* **681,** 897–904 (2008).
86. Bildfell, C., Hoekstra, H., Babul, A. & Mahdavi, A. Resurrecting the red from the dead: optical properties of BCGs in X-ray luminous clusters. *Mon. Not. R. Astron. Soc.* **389,** 1637–1654 (2008).
87. Best, P. N. *et al.* The host galaxies of radio-loud active galactic nuclei: mass dependences, gas cooling and active galactic nuclei feedback. *Mon. Not. R. Astron. Soc.* **362,** 25–40 (2005).
88. Best, P. N., Kaiser, C. M., Heckman, T. M. & Kauffmann, G. AGN-controlled cooling in elliptical galaxies. *Mon. Not. R. Astron. Soc.* **368,** 67L–71L (2007).
89. Dekel, A. & Birnboim, Y. Gravitational quenching in massive galaxies and clusters by clumpy accretion. *Mon. Not. R. Astron. Soc.* **383,** 119–138 (2008).
90. Khochfar, S. & Ostriker, J. P. O. Adding environmental gas physics to the semianalytic method for galaxy formation: gravitational heating. *Astrophys. J.* **680,** 54–69 (2008).
91. Naab, T., Johansson, P. H., Ostriker, J. P. & Efstathiou, G. Formation of early-type galaxies from cosmological initial conditions. *Astrophys. J.* **658,** 710–720 (2007).
92. Baldry, I. K. *et al.* Quantifying the bimodal color magnitude distribution of galaxies. *Astrophys. J.* **600,** 681–694 (2004).
93. Faber, S. M. *et al.* The centers of early-type galaxies with HST. IV. Central parameter relations. *Astron. J.* **114,** 1771–1796 (1997).
94. Naab, T., Khochfar, S. & Burkert, A. Properties of early type, dry galaxy mergers and the origin of massive elliptical galaxies. *Astrophys. J.* **636,** 81–84 (2006).
95. Cox, T. J. *et al.* The kinematic structure of merger remnants. *Astrophys. J.* **650,** 791–811 (2006).
96. Faber, S. M. *et al.* Galaxy luminosity functions to z∼1 from DEEP2 and COMBO-17: implications for red galaxy formation. *Astrophys. J.* **665,** 265–294 (2007).
97. Khalatyan, A. *et al.* Is AGN feedback necessary to form red elliptical galaxies? *Mon. Not. R. Astron. Soc.* **387,** 13–30 (2008).
98. Donahue, M., Horner, D. J., Cavagnolo, K. W. & Voit, G. M. Entropy profiles in the cores of cooling flow clusters of galaxies. *Astrophys. J.* **643,** 730–750 (2006).

Acknowledgements A.C. thanks his wife A. Fylaktou for assistance in making the Review readable for a broader audience.

Author Contributions A.C. initiated the project, wrote the first draft, and had editorial control throughout. He chose to have many co-authors to show that the Review reflects consensus within the field. S.M.F. made a major contribution to the structure and content of the Review. J.B., A.D., J.K. and R.M. participated extensively in the writing of the manuscript, A.B., P.B., and M.B. contributed significantly to individual sections. A.K., A.M. and P.B. produced Figs 2, 3 and 4, respectively. The other authors contributed mainly by providing comments on drafts and by participating in scientific discussions in connection with: X-ray observations of galaxy clusters (A.C.F.), galaxy formation (C.S.F., M.S.), the interaction of radiation with the interstellar medium (H.N.), the interaction of quasar winds with the interstellar medium (J.S.), and quasar winds and host galaxies (L.W.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence should be addressed to A.C. (acattaneo@aip.de).

ARTICLES

Adaptive prediction of environmental changes by microorganisms

Amir Mitchell[1], Gal H. Romano[2], Bella Groisman[1], Avihu Yona[1], Erez Dekel[3], Martin Kupiec[2], Orna Dahan[1]* & Yitzhak Pilpel[1,4]*

Natural habitats of some microorganisms may fluctuate erratically, whereas others, which are more predictable, offer the opportunity to prepare in advance for the next environmental change. In analogy to classical Pavlovian conditioning, microorganisms may have evolved to anticipate environmental stimuli by adapting to their temporal order of appearance. Here we present evidence for environmental change anticipation in two model microorganisms, *Escherichia coli* and *Saccharomyces cerevisiae*. We show that anticipation is an adaptive trait, because pre-exposure to the stimulus that typically appears early in the ecology improves the organism's fitness when encountered with a second stimulus. Additionally, we observe loss of the conditioned response in *E. coli* strains that were repeatedly exposed in a laboratory evolution experiment only to the first stimulus. Focusing on the molecular level reveals that the natural temporal order of stimuli is embedded in the wiring of the regulatory network—early stimuli pre-induce genes that would be needed for later ones, yet later stimuli only induce genes needed to cope with them. Our work indicates that environmental anticipation is an adaptive trait that was repeatedly selected for during evolution and thus may be ubiquitous in biology.

Microorganisms are constantly faced with environmental stimuli and stresses. The cellular response to such challenges has been intensively studied in several model organisms[1–4]. The simplest response strategy to a stimulus is to monitor the environment and to respond directly to it using designated mechanisms (Fig. 1). The environmental stress response in yeast represents a more complicated strategy in which the responses to many stresses are partially overlapping[1,2]. Theoretical work has shown that when a population of microorganisms evolves under erratic environmental fluctuations, cells may not effectively monitor the environment, but rather use stochasticity to randomly alternate between potential states[5] (Fig. 1). Stochastic switching might thus ensure that a portion of the population is prepared in advance for the unpredicted challenge[6,7]. However, other, more predictable, environments offer organisms the opportunity to adopt an alternative regulation strategy of anticipating an environmental change based on a preceding signal. The capacity of some complex multicellular eukaryotes to capture the statistics that govern the temporal connection between events in their environment, known as classical Pavlovian conditioning, serves as a central paradigm in the study of learning[8]. Here we ask whether genetic regulatory networks of microorganisms adaptively evolved to capture the temporal connections between subsequent stimuli in their environment. Most recently, 'anticipatory regulation' was discovered[9] (Fig. 1)—an association between environmental changes in bacteria. Specifically, this study investigated the response of *E. coli* to temperature increase that is followed by a drop in oxygen availability upon its entry to the digestive tract. Interestingly, these two signals show a symmetrical associative regulation pattern—each signal affects the expression of genes needed to cope with both (Fig. 1). Remarkably, the authors successfully decoupled the two responses during a laboratory evolution experiment in which the two signals were presented out-of-phase from one another. The ability to decouple the two responses is an indication that the coupling seen in the wild type is not a trivial combined response to the two stresses.

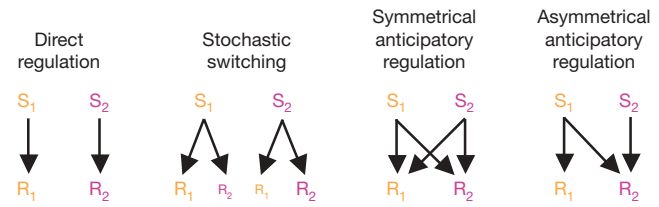


Figure 1 | Four possible regulation strategies in response to environmental stimuli. Under direct regulation, each of the stimuli, S_1 and S_2, activates exclusively the responses R_1 and R_2, respectively. Under stochastic switching, cells randomly sample either R_1 or R_2 in response to either S_1 or S_2. Under symmetrical anticipatory regulation, each of the stimuli activates both responses. Under asymmetrical anticipatory regulation, the stimulus that usually appears first in the ecology activates both responses, whereas the stimulus that appears later induces only the second response.

Here we show that biological systems that react to a unidirectional temporal order of environmental changes may manifest a more elaborate predictive capacity. This capacity is reflected in a corresponding asymmetric response strategy between subsequent stimuli, denoted S_1 and S_2 hereafter, and their designated responses R_1 and R_2, respectively. The first stimulus, S_1, activates both responses, R_1 and R_2, yet because the second stimulus, S_2, does not predict the appearance of S_1, it only activates its own response (Fig. 1). We propose three criteria to determine whether the observed cross-regulation pattern forms an adaptive anticipatory response strategy that could be selected for by evolution. First, asymmetric fitness advantage: pre-exposure to S_1 increases the fitness under S_2, yet pre-exposure to S_2 should not enhance fitness upon subsequent growth on S_1. This ensures that the natural order of stimuli was captured during evolution. Second,

[1]Department of Molecular Genetics, Weizmann Institute of Science Rehovot 76100, Israel. [2]Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv 69978, Israel. [3]Department of Molecular Cell Biology, Weizmann Institute of Science Rehovot 76100, Israel. [4]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

cost of preparation: pre-induction of genes needed to cope with S_2 is costly and not beneficial during S_1. This indicates that early induction is preserved due to a future benefit that is expected to exceed the cost, upon encounter with S_2. Third, specificity: the conditioned response is specific to S_1 and not to other unrelated stimuli, suggesting that anticipation evolved in response to the specific conditions of the ecological niche.

E. coli carbon source switch

During its life cycle *E. coli* alternates between two principal habitats, intestines of mammals and water, sediment and soil[10] (Supplementary Information). Focusing on the intestinal ecology reveals a predictable metabolic environment. Specifically, during passage along the digestive tract, exposure to lactose precedes exposure to another sugar, maltose[11]. We thus expect that this environment can select for asymmetric anticipation, so that bacteria that link the presence of lactose with future exposure to maltose are expected to activate the maltose genes already upon encounter with lactose. Such pre-induction is expected to allow cells to better use maltose upon its encounter.

We start by examination of anticipatory behaviour at the molecular level. To test systematically the promoter activity of all relevant operons in the presence of maltose and lactose we used an *E. coli* promoter-fused green fluorescent protein (GFP) library previously described[12]. Our experiments revealed the unidirectional regulation pattern expected under a conditioned response strategy (Fig. 2a and Supplementary Figs 1–4). Maltose operons, which are induced by maltose, are also induced, but to a lower level, by lactose. We note that the basic responsiveness of some of the maltose operons to lactose is not restricted only to the *E. coli* strain used in our study[13]. In contrast, and as expected by our hypothesis, the lactose operon shows no response to maltose (Fig. 2a and Supplementary Figs 1–4).

The crucial question is whether the observed asymmetric response strategy that implies anticipation is indeed adaptive—that is, whether it could be selected for during evolution. Following the criteria outlined previously, we first validated that the conditioned response indeed provides an asymmetric fitness advantage in an alternating sugar environment. We monitored cell growth on maltose after an initial growth phase on lactose. Reassuringly, we observed that the wild-type strain displays a fitness advantage if growth on maltose was preceded by growth on lactose (Fig. 3). We ensured that this advantage does not originate from the metabolic contribution of lactose as a carbon source (Supplementary Information). In addition, and as required by our first principle, we found that pre-exposure to maltose before growth on lactose does not improve fitness (Fig. 3).

Our second criterion for adaptiveness requires that pre-induction of the maltose genes is maladaptive or neutral during growth on lactose. This criterion is crucial to rule out the possibility that the maltose genes are induced during exposure to lactose simply because they are needed during that phase. Built into the classical conditioning paradigm from the cognitive context is the possible extinction of the association. To put in present case terms, it is expected that repeated exposure to lactose without consequent arrival of maltose would select for weakening of the conditioned response due to the futile cost of preparation. Accordingly, we have examined laboratory-evolved strains of *E. coli*, which grew for 500 generations on high levels of lactose yet without exposure to maltose[14]. We measured the promoter activity of relevant operons in three lineages that evolved independently under various lactose concentrations. Notably, we found that, in all three lines, the maltose operons show almost no activity in response to lactose (Fig. 2b and Supplementary Fig. 5). Reassuringly, the strains seem to have preserved the ability to activate the maltose genes in response to maltose itself (Supplementary Fig. 6), indicating that only the asymmetrical cross talk between the two pathways was removed during this laboratory evolution period. These results imply that the conditioning observed in the wild-type strain is costly and that without a subsequent benefit that should exceed this cost, purifying selection acts to eliminate this cross talk. Consistent with the second criterion, the growth advantage

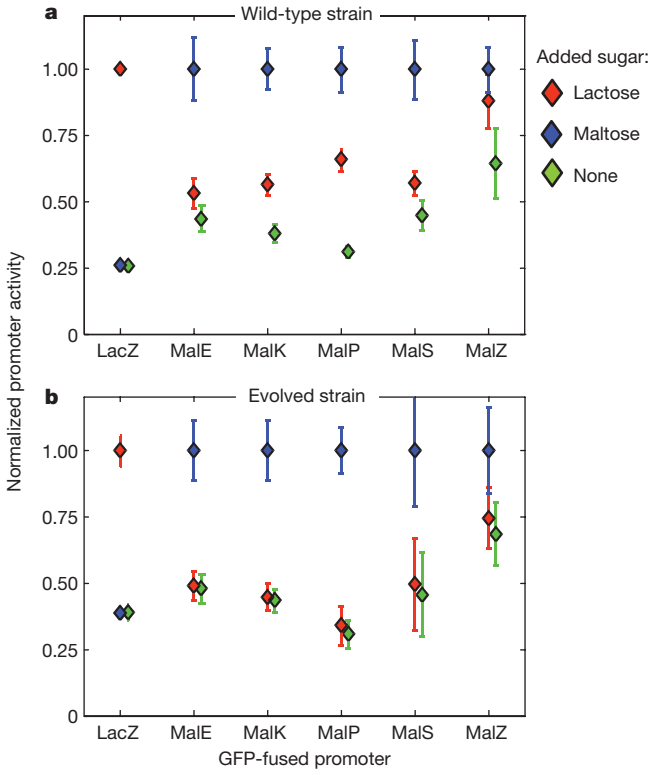


Figure 2 | Conditioned response in *E. coli* sugar metabolism. Promoter activity in wild type (**a**) and the evolved strain (**b**). Colours mark the added sugar. Error bars denote standard deviation of four repeats. Promoter activities of four out of five maltose operons in wild type are significantly higher under lactose relative to the untreated culture (*t*-test, $P < 0.01$ according to a Bonferroni adjustment). In contrast, none of the promoters show increased activity in the evolved strain. The experiment was done in M9-Glu medium and was repeated using glycerol as a background carbon source (Supplementary Fig. 3) and further verified using quantitative polymerase chain reaction with reverse transcription (RT–PCR) (Supplementary Figs 2 and 4).

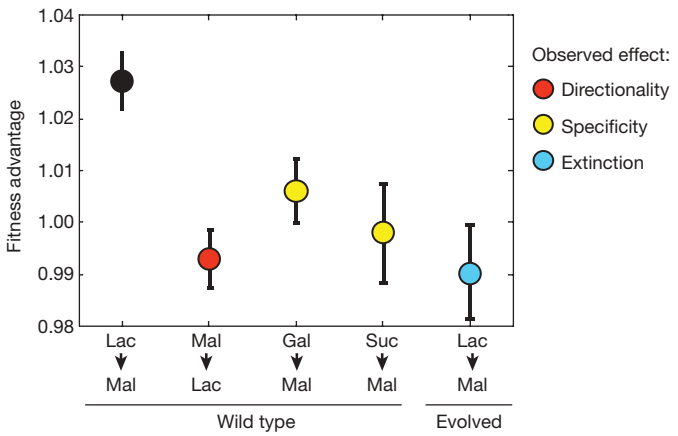


Figure 3 | Fitness in an alternating sugar environment. Two isogenic populations were grown on M9-Gly either supplemented with a sugar as S_1 (treated) or not (untreated). An equal amount of cells was diluted into media containing the second sugar (S_2) as a sole carbon source. The ratio treated/untreated (after 2 h) represents the fitness advantage originating from pre-exposure to S_1. The error bars denote standard deviation of three repeats. All ratios are normalized to the ratio measured in a control experiment in which S_2 was glycerol (Methods). A significant fitness advantage was observed only in the transfer of wild-type cells from lactose to maltose ($P = 0.02$, *t*-test).

observed in the wild type due to exposure to lactose before maltose is lost in the evolved strain that features extinction (Fig. 3).

Finally, as required by the specificity criterion, we ensured that pre-exposure to other sugars does not improve growth of cells upon transfer to maltose. As can be seen in Fig. 3, two alternative related carbon sources, galactose and sucrose, cannot substitute lactose as a preceding signal before maltose. This indicates that the observed anticipation evolved in response to the specific stimuli pair that is presented to the organism in its ecological niche. In summary, we conclude that anticipation of subsequent carbon source changes is an adaptive trait that has been selected by the natural ecology of *E. coli*.

The shift from fermentation to respiration in yeast

We next proceeded to a more complicated case that may also manifest environmental change anticipation. We examined how environmental anticipation is encoded in the regulatory network of *S. cerevisiae*, and whether this response strategy has been selected for because of a net fitness advantage it entails. We examined conditions that emulate some of the stresses that occur during the process of alcoholic brew production, such as wine, by yeast[15]: Fig. 4a describes this process, which starts at a highly osmotic and low pH environment, followed by potentially lethal temperature raise due to vigorous fermentation and accumulation of ethanol, culminating in the diauxic shift—the exhaustion of fermentable sugars (for example, glucose) and the switch to oxidative respiration[16–19]. Respiration is challenging to the cell because it eventually generates oxygen radicals[20]. We checked whether yeast cells can cope better with these stresses provided that they appear in their natural temporal order, that is, whether yeasts use early stresses as predictive signals for the likely arrival of later stresses. Such behaviour would indicate that yeast cells are preparing for later challenges even before they arise.

To address these questions we systematically studied the survival of *S. cerevisiae* under stresses that comprise the switch from its preferred catabolic state, fermentation, to the alternative one, respiration (Fig. 4). As a potential manifestation of anticipation, we focused on the cross-protection phenotype between different stress pairs[21,22]. A cross-protection phenotype between two stresses exists if pre-exposure to one stress improves the survival of cells under a subsequent stress. Under our hypothesis we expect an asymmetrical cross-protection phenotype between stress pairs. Particularly, we predict that directionality of cross-protection will be in accordance with the order of events during the transition from fermentation to respiration. For example, because heat shock and ethanol accumulation precede oxidative stress in the ecology, we expect improved survival under oxidative stress if it is preceded by either heat shock or ethanol stress, yet we do not expect cross-protection if the stresses are applied in the reverse order. Figure 4a shows the measured protection phenotype systematically tested for stresses that occur before and after the diauxic shift. In accordance with our hypothesis, we observe that the cross-protection phenotype is indeed extremely asymmetric and, most importantly, that the directionality of protection fits the natural temporal context. Specifically, for each stress pair protection is always stronger in the direction found in the natural environment, thus fulfilling the asymmetric fitness advantage criterion. In a more faithful emulation of the natural habitat we have also applied the three stresses, heat shock, ethanol and oxidative stress, sequentially and observed an equally high cross-protection activity (Supplementary Information). Note that activation of the

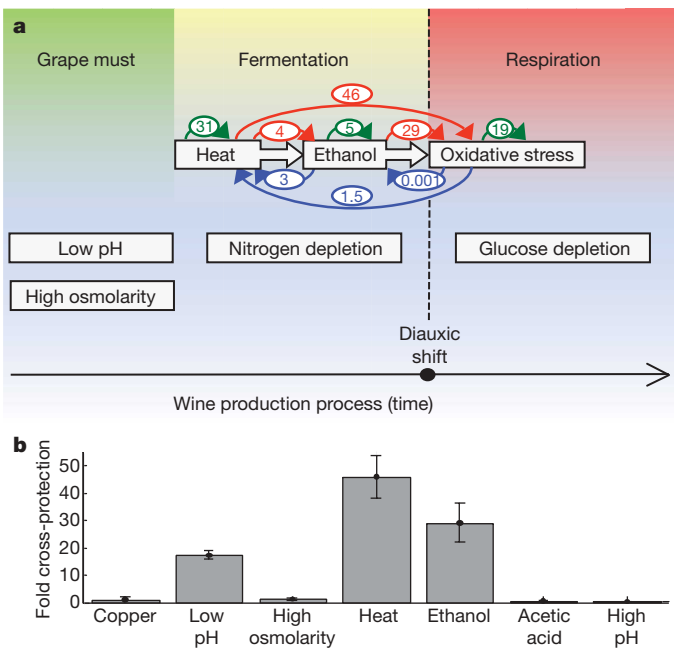


Figure 4 | Cross-protection in the context of the diauxic shift. **a**, Stress sequence in the process of wine production and measured cross-protection phenotype. 'Must' is the medium in the beginning of the wine production process. The values denote the fold protection gained by pre-exposure to a mild stress followed by a severe stress. Red and blue arrows mark cross-protection when stress order is in accordance with the ecology or in reverse order, respectively. Green arrows denote auto-protection. **b**, Specificity of cross-protection against oxidative stress. The columns mark the observed fold cross-protection provided by exposure to mild stress before severe oxidative stress. Error bars donate standard deviations of three repeats.

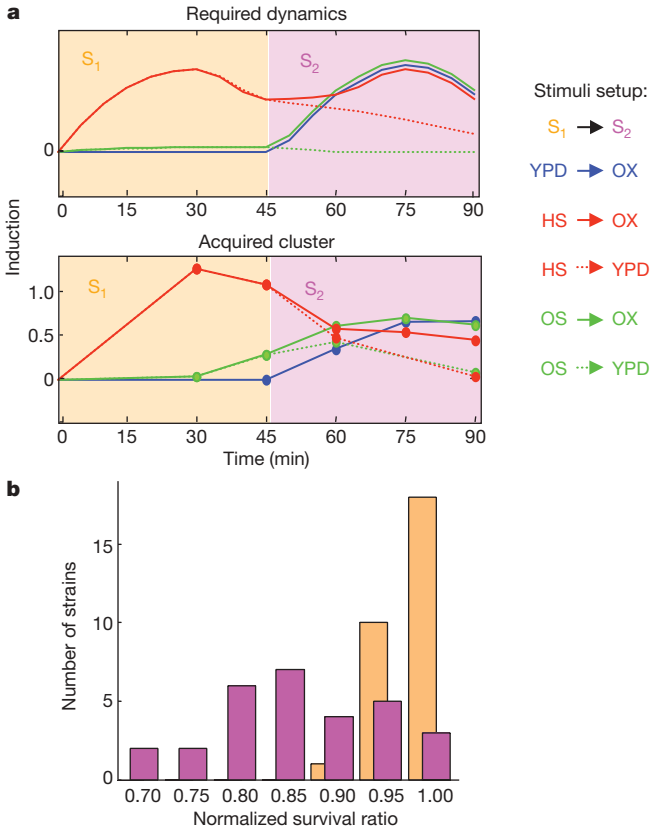


Figure 5 | Candidate genes underlying the asymmetrical protection between heat and oxidative stresses. **a**, Genome-wide expression was measured in response to sequential stress pairs (untreated, YPD; osmotic stress, OS; heat shock, HS; oxidative stress, OX). The top panel marks the expression dynamics of an ideal gene that may facilitate cross-protection against oxidative stress. The bottom panel marks the mean profile of 300 genes matching the desired pattern (Methods). **b**, A histogram of sensitivity values of 29 deletion strains to heat (orange) and oxidative (purple) stresses. The survival of each strain was normalized to the survival of the wild type to test for increased sensitivity (individual strain sensitivity is shown in Supplementary Fig. 7 and Supplementary Table 5).

environmental stress response[1,2], namely a set of genes that are transcribed in response to a wide spectrum of stresses, is less likely to explain the observed cross-protection activity. Whereas the environmental stress response is more likely to result in a symmetrical protection, we observe a very asymmetric effect. Furthermore, and as required by the third criterion, we observe that protection against oxidative stress is specific to stresses that precede respiration in the natural environment, whereas stresses unrelated to the process do not provide cross-protection (Fig. 4b).

We next investigated a potential molecular mechanism that might account for the observed cross-protection capacity. We focused on the pair of stresses that exhibits the strongest cross-protection phenotype, heat shock and oxidative stress, and measured genome-wide gene expression in response to these stresses when they were introduced in isolation or one after the other. In particular, we applied a heat shock as the first stress (S_1), followed by an oxidative stress (S_2). As a control we also examined the response to oxidative stress that is preceded by an osmotic stress, a condition we found not to cross-protect significantly against oxidative stress (Fig. 4b). We used these series of events to define an ideal expression profile of genes that can underlie the cross-protection phenotype and ultimately the anticipation capacity (Fig. 5a, see Methods for profile details). We also defined a control profile, using the control osmotic stress (instead of heat shock) as an early stimulus. Our search revealed a cluster of 300 genes with the desired profile (Fig. 5a). Consistent with asymmetric anticipation, this cluster shows a significant overlap with a set of genes that are annotated in the literature (Proteome database) as essential in oxidative stress but not in heat shock ($P = 0.013$). However, this cluster does not overlap significantly with a set of genes that are annotated as essential in heat shock and not in oxidative stress ($P = 0.2$). Analysis of functional enrichment using Gene Ontology annotations[23] revealed a similar trend—a significant enrichment of oxidative stress categories (Supplementary Table 3), such as oxidoreductase activity, and many mitochondria-related categories, consistent with a known role of mitochondria for coping with oxidative stress in yeasts[24]. However, no heat-shock-related categories were found to be enriched in this cluster. Reassuringly, the genes that corresponded to the control profile (in which we still used oxidative stress as an S_2, yet with osmotic stress as an unprotective S_1) did not yield similar enrichment for oxidative-related activities (Supplementary Table 4). This indicates that whereas an S_1 stimulus that provides cross-protection increases in advance the expression level of genes relevant for the R_2 response, a control S_1 signal does not affect these R_2 genes.

We then tested whether induction of the conditioned gene set is neutral in heat shock, as required by the second criterion. Towards this aim we collected 29 strains, each of which is deleted for one of the genes from the cluster described above, and tested their sensitivity to heat shock and oxidative stress (Methods). Analysis of survival ratios reveals that whereas 65% of the strains show increased sensitivity to oxidative stress, none shows a significantly increased sensitivity to heat shock relative to the wild-type strain (Fig. 5b and Supplementary Fig. 7). This is a direct indication that the expression of many of the heat-induced genes is at least neutral, if not maladaptive, under heat shock. Taken together with the fact that these genes are essential under oxidative stress, this experiment corroborates the claim that an early preparation to oxidative stress during heat shock is adaptive in this system.

Future outlook

An open question is how prevalent anticipatory regulation is in the world of microorganisms. One way to address this issue is to develop a mathematical framework that estimates the potential advantage of a conditioned response in a given ecology. Intuitively, the added value from anticipation depends both on environmental and on biological parameters. Two cellular parameters, gain and cost, should be considered. A conditioned response is beneficial provided that benefit gained from anticipation exceeds the cost of early preparation. These parameters are in turn dependent on the typical time constants of the environment, for example, the time gap between the appearances of the two stimuli. Additionally, predictable environments in which the stimuli are often sequentially coupled promote anticipation. A simple equation captures the relationship between these variables and their effect on the net fitness:

$$\Delta F = p(B(\Delta t) - C(\Delta t)) - (1 - p)(C(t_{S1})),$$

where ΔF is the difference in fitness between an anticipating cell and a cell that adheres to direct regulation (Fig. 1), p is the probability that S_2 will occur given that S_1 occurred, and B and C are functions describing the benefit and cost of early preparation, respectively. Δt is the length of the time interval between the appearance of S_1 and S_2, and t_{S1} is the duration of the first stimulus. The dependence of fitness on the interplay between the key factors is shown in Supplementary Fig. 8. We note that quantitative predictions from the above phenomenological equation can be drawn by incorporating into the model biologically meaningful and realistic parameters, such as measurable gain, cost and timescales (A.M. *et al.* in preparation).

To conclude, as we observed that anticipation at the cellular level appears to be adaptive both in prokaryotes (*E. coli*) and eukaryotes (*S. cerevisiae*), and under remote systems such as sugar metabolism and stress response, we propose that it may be prevalent in a diverse range of species. Asymmetric anticipation, along with associative learning[9], memory[25] and physiological adaptation[26], represent a powerful set of tools used by microorganisms to achieve a sophisticated environmental response.

METHODS SUMMARY

Strains. *E. coli* MG1655 was used for most experiments. An *E. coli* GFP reporter library[12] was used to monitor operon expression. The evolved *E. coli* strains were described previously[14]

S. cerevisiae BY4741 (*MATa*; *his3Δ1*; *leu2Δ0*; *met15Δ0*; *ura3Δ0*) was used in all yeast experiments. All deletion strains were obtained from the *Saccharomyces* Genome Deletion Project[27].

***S. cerevisiae* cross-protection.** An overnight culture was diluted into fresh YPD medium and grown to a concentration of 2×10^6 cells per ml. Cells were diluted 1:2 into the first stress (S_1). Transfer into the second stress (S_2) was done by adding treated medium after centrifugation. Stresses were calibrated to achieve a mild effect (~50% survival) for S_1 and a severe effect (~0.5% survival) for S_2 (Supplementary Table 1). Samples were taken from cultures at the end of each treatment and plated. Each experiment was carried out in three repeats. The fold protection was calculated as described in Supplementary Equation (1).

***E. coli* fitness advantage.** Overnight cultures were diluted into M9-Gly plus S_1 (for example, lactose; treated) or M9-Gly (untreated). After 3 h, population size was determined and used to dilute (~1:100) an equal amount of cells into a new growth medium containing low levels of S_2 (for example, maltose) as a sole carbon source. To account for a potential metabolic effect of residual S_1, an identical trace amount of S_1 was added to the untreated growth medium. The ratio treated/untreated after 2 h of growth was used as an indication of fitness advantage originating from pre-exposure to S_1. To rule out the possibility that the growth advantage observed due to early exposure to lactose originates from its metabolic value rather than its rule as a conditioning signal, we preformed a control experiment (Supplementary Information).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 March; accepted 7 May 2009.
Published online 17 June 2009.

1. Causton, H. C. *et al.* Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* **12**, 323–337 (2001).
2. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
3. Cheung, K. J. *et al.* A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. *Genome Res.* **13**, 206–215 (2003).
4. Tirosh, I., Weinberger, A., Carmi, M. & Barkai, N. A genetic signature of interspecies variations in gene expression. *Nature Genet.* **38**, 830–834 (2006).

5. Kussell, E. & Leibler, S. Phenotypic diversity, population growth, and information in fluctuating environments. *Science* **309,** 2075–2078 (2005).
6. Acar, M., Mettetal, J. T. & van Oudenaarden, A. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genet.* **40,** 471–475 (2008).
7. Balaban, N. Q. *et al.* Bacterial persistence as a phenotypic switch. *Science* **305,** 1622–1625 (2004).
8. Pavlov, I. P. *Conditioned Reflexes* (Oxford Univ. Press, 1927).
9. Tagkopoulos, I., Liu, Y. C. & Tavazoie, S. Predictive behavior within microbial genetic networks. *Science* **320,** 1313–1317 (2008).
10. Savageau, M. A. *Escherichia coli* habitats, cell-types, and molecular mechanisms of gene control. *Am. Nat.* **122,** 732–744 (1983).
11. Savageau, M. A. Demand theory of gene regulation. II. Quantitative application to the lactose and maltose operons of *Escherichia coli. Genetics* **149,** 1677–1691 (1998).
12. Zaslaver, A. *et al.* A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli. Nature Methods* **3,** 623–628 (2006).
13. Soupene, E. *et al.* Physiological studies of *Escherichia coli* strain MG1655: Growth defects and apparent cross-regulation of gene expression. *J. Bacteriol.* **185,** 5611–5626 (2003).
14. Oxman, E., Alon, U. & Dekel, E. Defined order of evolutionary adaptations: experimental evidence. *Evolution* **62,** 1547–1554 (2008).
15. Mortimer, R. K. Evolution and variation of the yeast (*Saccharomyces*) genome. *Genome Res.* **10,** 403–409 (2000).
16. Sablayrolles, J. M. & Barre, P. Kinetics of alcoholic fermentation under anisothermal enological conditions. *Am. J. Enol. Vitic.* **44,** 127–133 (1993).
17. Johnston, M. & Carlson, M. in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression* (eds Jones, E. W., Pringle, J. R. & Broach, J. R.) 193–282 (Cold Spring Harbor Laboratory, 1992).
18. Bauer, F. F. & Pretorius, I. S. Yeast stress response and fermentation efficiency: How to survive the making of wine - a review. *SAJEV* **21,** 27–51 (2000).
19. Pretorius, I. S. Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. *Yeast* **16,** 675–729 (2000).
20. Maris, A. F. *et al.* Diauxic shift-induced stress resistance against hydroperoxides in *Saccharomyces cerevisiae* is not an adaptive stress response and does not depend on functional mitochondria. *Curr. Genet.* **39,** 137–149 (2001).
21. Sanchez, Y., Taulien, J., Borkovich, K. A. & Lindquist, S. Hsp104 is required for tolerance to many forms of stress. *EMBO J.* **11,** 2357–2364 (1992).
22. Berry, D. B. & Gasch, A. P. Stress-activated genomic expression changes serve a preparative role for impending stress in yeast. *Mol. Biol. Cell* **19,** 4580–4587 (2008).
23. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32** (Database issue), D258–D261 (2004).
24. Grant, C. M., MacIver, F. H. & Dawes, I. W. Mitochondrial function is required for resistance to oxidative stress in the yeast *Saccharomyces cerevisiae. FEBS Lett.* **410,** 219–222 (1997).
25. Zacharioudakis, I., Gligoris, T. & Tzamarias, D. A yeast catabolic enzyme controls transcriptional memory. *Curr. Biol.* **17,** 2041–2046 (2007).
26. Braun, E. & Brenner, N. Transient responses and adaptation to steady state in a eukaryotic gene regulation system. *Phys. Biol.* **1,** 67–76 (2004).
27. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285,** 901–906 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Trattner–Frenkel and Z. Bloom for their help and support in the yeast microarray experiments. We thank members of the Pilpel laboratory for many discussions. We thank E. Schneidman, E. Ben-Jacob, M. Springer, A. Tanay, U. Alon and D. Cavalieri for discussions and advice. We thank U. Alon for providing the promoter–GFP fused plasmids. We thank the Tauber Foundation, the Minerva Foundation, the Israel Science Foundation 'Bikura program', the European Research Council 'Ideas Program' and the Ben May Foundation for grant support. M.K. was supported from grants from the Israel Science Foundation and the Israeli Ministry of Science and Technology.

Author Contributions A.M. raised the original idea and performed all the experiments; G.R., B.G. and A.Y. participated in experiments; E.D. evolved the *E. coli* strain; A.M., O.D. and Y.P. designed the experiments; A.M., M.K., O.D. and Y.P. analysed the data; O.D. and Y.P. supervised the project; A.M., O.D. and Y.P. interpreted the results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.P. (pilpel@weizmann.ac.il).

METHODS

***S. cerevisiae* media.** All experiments were carried out in YPD medium (2% yeast extract, 1% peptone, 1% dextrose) at 30 °C.

Sensitivity of deleted *S. cerevisiae* strains to heat and oxidative stress. To examine the importance of the conditioned gene cluster under heat and oxidative stresses, 29 genes were chosen for additional experiments. This subset of genes exhibited a two-phase induction profile: initial induction under heat shock that further increased under oxidative stress. We then used strains in which each of the 29 genes was deleted to check for increased stress sensitivity as follows.

Cultures were grown to stationary phase in a 96-well plate, diluted (1:20) into fresh YPD medium and grown for additional 3 h. Cells were then diluted into either mild heat shock (37 °C for 30 min) or mild oxidative stress (H_2O_2 1 mM for 30 min). The stresses were calibrated to achieve 90% survival in the wild-type strain. As a control, untreated cultures were diluted into YPD. All cultures were then diluted (1:20) into fresh YPD and grown for 6 h. Population size was monitored using a multi-well spectrophotometer at 595 nm. Survival ratio was calculated by dividing the optical density of stress-treated cultures by the optical density of untreated cultures. A *t*-test was used to determine whether the difference in the survival ratio of a deletion strain and the wild type is statistically significant in a given stress.

Genome-wide expression experiments. An overnight culture was diluted into fresh YPD and grown to concentration of 2.5×10^7 cells per ml. Cells were diluted 3:4 into a fresh medium containing S_1 (KCl 0.8 M, heat shock 40 °C or YPD) and grown for 45 min. Cells were then diluted 1:10 into S_2 (H_2O_2 0.66 mM or YPD). Stress levels were calibrated to have relatively minor effects on survival. Aliquots were removed 0, 30 and 45 min after the addition of S_1 whereas in S_2 aliquots were removed after 15, 30 and 45 min. RNA was extracted using MasterPure, followed by hybridization to Affymetrix yeast 2.0 microarrays. The expression data set was deposited in the GEO database (GSE15936).

Identifying candidate genes that facilitate conditioning. We defined an ideal expression profile of a gene that can facilitate the observed cross-protection phenotype (heat shock but not osmotic stress protects against oxidative stress). The profile is defined accordingly: (1) induction in response to heat shock is 0.2 or higher in $\log_2$ scale; (2) induction in response to oxidative stress is 0.2 or higher in $\log_2$ scale; (3) induction under osmotic stress is significantly lower than in heat shock, 0.4 in $\log_2$ scale; (4) induction is maintained higher in the transfer from heat shock to oxidative stress relative to transferring from heat shock to YPD.

A control profile is similarly defined by using osmotic stress instead of heat shock in the above requirements. The results are qualitatively similar under alternative cut-off values (not shown).

***E. coli* media.** All experiments were done in M9 medium (1 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.05% casamino acids and 5 ng ml^{-1} thiamine), supplemented with the appropriate carbon source.

Three basic media were used: M9, M9-Gly (0.1% glycerol) and M9-Glu (0.1% glucose plus 20 mM cAMP; cAMP was added to avoid glucose repression[28]).

To test the effect of various treatments the media were supplemented with the following: 10 mM lactose, 5 mM maltose (these concentrations allow maximal growth), 10 mM galactose, 10 mM sucrose (these concentrations are equal to the concentration of lactose), 0.15 mM IPTG (this concentration achieves saturation of the lactose operon induction[29]) and 25 μg ml^{-1} kanamycin.

Monitoring operon transcription using the promtor-fused GFP library. Overnight cultures, each carrying a unique plasmid with a specific promoter fused to GFP, were diluted into fresh M9-Glu media, grown for 1 h at 37 °C and then treated with sugars. Expression and cell growth (at 30 °C) were monitored simultaneously using a multi-well reader (fluorescence at 495/520 nm, optical density at 595 nm). The normalized GFP level and promoter activity were calculated similarly to in ref. 28. Briefly, the GFP signal was calculated after subtraction of the medium fluorescence and cell autofluorescence. GFP per cell was calculated dividing the GFP by the optical density. Promoter activity ((dGFP/dt)/optical density) was taken as the average promoter activity measured in a time window of 1 h of exponential growth.

The following operons were examined in our research: MalEFG, MalK-lamB-malM, MalPQ, MalS, MalZ, LacZYA (referred to by the name of the first gene). MalT was not included in our analysis because it shows no responsiveness to maltose.

Monitoring operon transcription using the quantitative RT–PCR. Overnight cultures were diluted into either M9-Gly or M9-Glu, alone or supplemented with either lactose or maltose, and grown until the logarithmic phase was reached. RNA was extracted using RNeasy Mini kit and used as a template for quantitative RT–PCR (LightCycler 480 system).

28. Kaplan, S. *et al.* Diverse two-dimensional input functions control bacterial sugar genes. *Mol. Cell* **29,** 786–792 (2008).
29. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436,** 588–592 (2005).

ARTICLES

CCR3 is a target for age-related macular degeneration diagnosis and therapy

Atsunobu Takeda[1]*, Judit Z. Baffi[1]*, Mark E. Kleinman[1]*, Won Gil Cho[1]*, Miho Nozaki[1,3], Kiyoshi Yamada[1], Hiroki Kaneko[1], Romulo J. C. Albuquerque[1,2], Sami Dridi[1], Kuniharu Saito[1], Brian J. Raisler[1,2], Steven J. Budd[4], Pete Geisen[4], Ariel Munitz[5], Balamurali K. Ambati[6,7], Martha G. Green[1], Tatsuro Ishibashi[8], John D. Wright[4], Alison A. Humbles[9]†, Craig J. Gerard[9], Yuichiro Ogura[3], Yuzhen Pan[10], Justine R. Smith[10], Salvatore Grisanti[11], M. Elizabeth Hartnett[4], Marc E. Rothenberg[5] & Jayakrishna Ambati[1,2]

Age-related macular degeneration (AMD), a leading cause of blindness worldwide, is as prevalent as cancer in industrialized nations. Most blindness in AMD results from invasion of the retina by choroidal neovascularisation (CNV). Here we show that the eosinophil/mast cell chemokine receptor CCR3 is specifically expressed in choroidal neovascular endothelial cells in humans with AMD, and that despite the expression of its ligands eotaxin-1, -2 and -3, neither eosinophils nor mast cells are present in human CNV. Genetic or pharmacological targeting of CCR3 or eotaxins inhibited injury-induced CNV in mice. CNV suppression by CCR3 blockade was due to direct inhibition of endothelial cell proliferation, and was uncoupled from inflammation because it occurred in mice lacking eosinophils or mast cells, and was independent of macrophage and neutrophil recruitment. CCR3 blockade was more effective at reducing CNV than vascular endothelial growth factor A (VEGF-A) neutralization, which is in clinical use at present, and, unlike VEGF-A blockade, is not toxic to the mouse retina. *In vivo* imaging with CCR3-targeting quantum dots located spontaneous CNV invisible to standard fluorescein angiography in mice before retinal invasion. CCR3 targeting might reduce vision loss due to AMD through early detection and therapeutic angioinhibition.

AMD affects 30–50 million people globally, with approximately 90% of severe vision loss attributed to CNV[1]. The worldwide prevalence of CNV is expected to double in the next decade owing to population ageing. Targeting the pro-angiogenic cytokine VEGF-A has been validated in patients with CNV[2–4]. However, substantial improvement of vision only occurs in one-third of patients treated with VEGF-A antagonists, and one-sixth of treated patients still progress to legal blindness. Moreover, safety concerns about the continual blockade of VEGF-A, which is constitutively expressed in the normal adult human retina[5], are emerging[6,7]. Thus, treatment strategies on the basis of more specific targeting of CNV are desirable. However, no molecular marker specific for human CNV has yet been reported.

CCR3 expression restricted to CNV in human eyes

In our studies examining the role of chemokines in angiogenesis, we discovered that CCR3 (also known as CD193)—a chemokine receptor best known for its role in promoting eosinophil and mast cell trafficking[8]—was expressed in human choroidal endothelial cells (CECs) only in the context of CNV due to AMD, and not in other non-proliferating or proliferating choroidal vasculature (Fig. 1). Immunolocalization studies showed that CCR3 was expressed in CECs of all examined specimens of surgically excised choroidal neovascular tissue from patients with AMD (18 out of 18) who had not received prior AMD treatment (Fig. 1a, b and Supplementary Fig. 1). In contrast, CCR3 was not expressed in CECs in the choroid of any patients with early (atrophic) AMD (0 out of 10), or in age-matched patients without AMD (0 out of 10) (Fig. 1c, d). CCR3 was also not immunolocalized in surgically excised tissue from patients with epiretinal fibrotic membranes (0 out of 6), or in CECs in patients with choroidal melanoma (0 out of 8) (Fig. 1e, f). Collectively, these data point to a highly specific pattern of CCR3 expression ($P = 7 \times 10^{-14}$, exact contingency table test) in CECs in neovascular AMD. Furthermore, we identified the expression of the CCR3 ligands eotaxin-1 (also known as CCL11), -2 (CCL24), and -3 (CCL26) in all examined specimens of surgically excised choroidal neovascular tissue from patients with AMD who had not received previous AMD treatment (Fig. 1g–j), suggesting that the eotaxin–CCR3 axis could be involved in this disease state. Notably, despite the abundance of eotaxins, eosinophils and mast cells were not identified in human CNV (Supplementary Fig. 2), consistent with earlier findings[9].

CCR3 stimulation promotes CEC migration and proliferation

The best determined pathological function of CCR3 so far has been its role in allergic diseases, such as asthma[10–14] and eosinophilic esophagitis[15]. There is a single report of its direct role in angiogenesis[16]. Although eosinophils and mast cells have been reported to be

[1]Department of Ophthalmology & Visual Science, [2]Department of Physiology, University of Kentucky, Lexington, Kentucky 40506, USA. [3]Department of Ophthalmology and Visual Science, Nagoya City University Graduate School of Medical Sciences, Nagoya 467-8601, Japan. [4]Department of Ophthalmology, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [5]Division of Allergy and Immunology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, Ohio 45229, USA. [6]Department of Ophthalmology and Visual Sciences, Moran Eye Center, University of Utah School of Medicine, Salt Lake City, Utah 84132, USA. [7]Department of Ophthalmology, Veterans Affairs Salt Lake City Healthcare System, Salt Lake City, Utah 84148, USA. [8]Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University, Fukuoka 812-8582, Japan. [9]Department of Medicine, Children's Hospital, Harvard Medical School, Boston, Massachusetts 02215, USA. [10]Casey Eye Institute, Oregon Health and Science University, Portland, Oregon 97239, USA. [11]Department of Ophthalmology, University of Luebeck, D-23538 Lübeck, Germany. †Present address: Respiratory, Inflammation and Autoimmunity, Medimmune, Inc., Gaithersburg, Maryland 20878, USA.
*These authors contributed equally to this work.

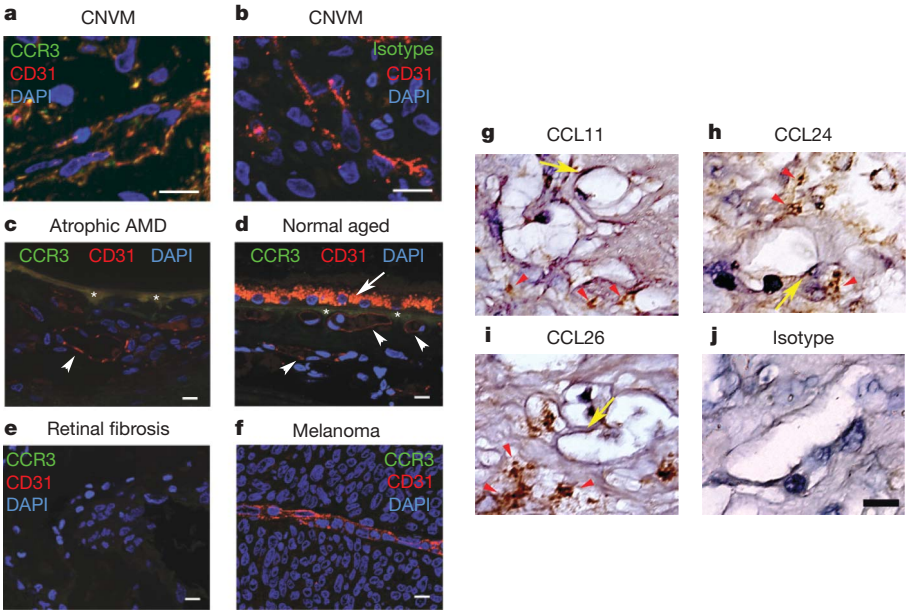


Figure 1 | CCR3 and eotaxins are expressed in CNV. **a**, **b**, Immunofluorescence shows that CCR3-receptor expression (green) colocalizes with $CD31^+$-expressing (red) blood vessels in surgically excised human AMD choroidal neovascular tissue (CNVM). Nuclei were stained blue with 4,6-diamidino-2-phenylindole (DAPI). The specificity of CCR3 staining in **a** is confirmed by the absence of staining with an isotype control IgG (green) in **b**. Individual red and green fluorescence channels are shown in Supplementary Fig. 1. **c**, **d**, CCR3 is not immunolocalized in $CD31^+$ (red) blood vessels (white arrowheads) in the choroid of patients with atrophic AMD who do not have CNV (**c**), or in aged patients without AMD (**d**). Autofluorescence of RPE (white arrow) and Bruch's membrane (asterisks) overlying choroid are seen. **e**, **f**, CCR3 is not expressed in surgically excised avascular retinal fibrosis tissue (**e**) or in the blood vessel of choroidal melanoma (**f**). **g**–**j**, Immunohistochemistry (golden brown reaction product) shows expression of CCL11 (**g**), CCL24 (**h**), and CCL26 (**i**) in surgically excised AMD choroidal neovascular tissue, primarily in the stroma (red arrowheads) but also in the blood vessels (yellow arrows). The specificity of staining is confirmed by the absence of staining with isotype control IgG (**j**). Scale bars, 10 μm.

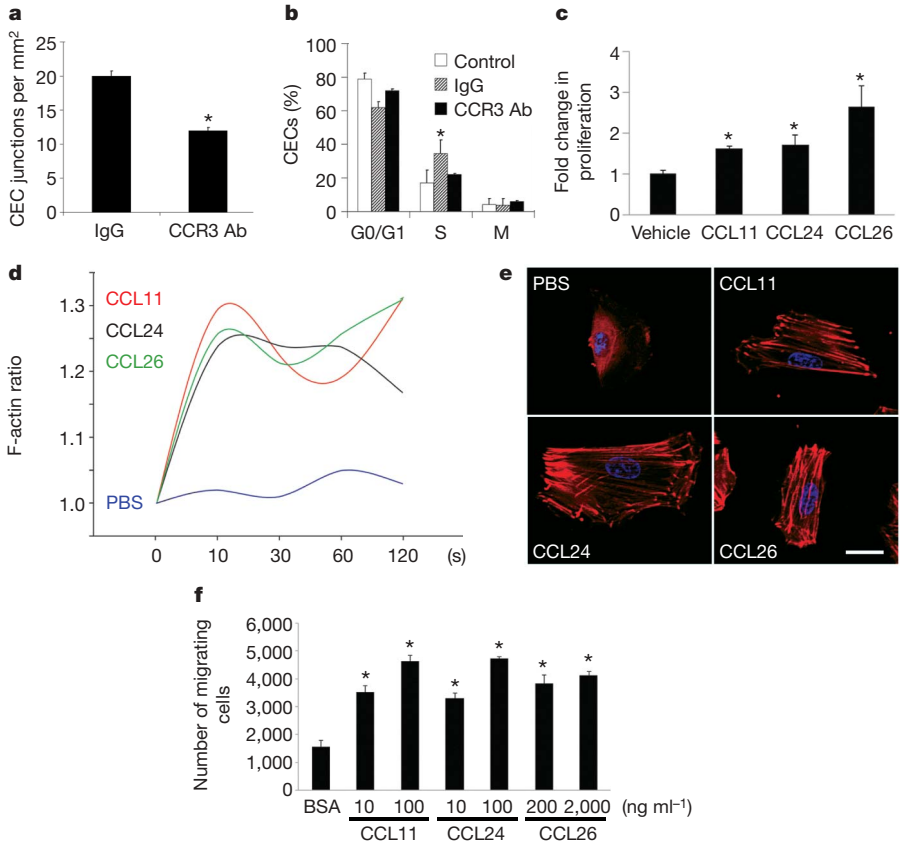


Figure 2 | CCR3 activation promotes angiogenesis. **a**, Tube formation of primary human CECs in Matrigel *in vitro* was reduced by neutralizing anti-human CCR3 antibodies (CCR3 Ab) compared to isotype IgG. $n = 6$, $^*P < 0.05$ compared to isotype IgG. **b**, Fraction of $CD31^+$ $VEGFR2^+$-gated mouse CECs *in vivo* in the proliferative state (S phase) was increased 5 days after laser injury in wild-type mouse eyes compared to control (uninjured eyes), and was reduced by intraocular administration of neutralizing anti-mouse CCR3 antibody compared to isotype IgG. $n = 6$–10, $^*P < 0.05$ compared to IgG treatment. **c**, Stimulation with eotaxins for 24 h induced human CEC proliferation. $n = 4$, $^*P < 0.05$ compared to BSA treatment. **d**, **e**, Stimulation with eotaxins, but not with PBS, induced actin polymerization in human CECs. The relative F-actin content is expressed as the ratio of the mean channel fluorescence between eotaxin-stimulated and media-alone-stimulated cells (**d**). Rhodamine–phalloidin staining (red) shows F-actin fibre formation in eotaxin-stimulated cells (**e**). Nuclei were stained blue with DAPI. Data representative of 3–4 independent experiments are shown. In **c** and **e**, 10 ng ml^{-1} CCL11, 100 ng ml^{-1} CCL24 and 2 μg ml^{-1} CCL26 were used. Scale bar, 20 μm. **f**, Stimulation with eotaxins for 16 h induces dose-dependent migration of human CECs across 8-μm pore size Transwells. $n = 5$–10, $^*P < 0.05$ compared to BSA treatment. Statistical significance was determined by Mann–Whitney U test (**a**–**c**, **f**); error bars depict s.e.m.

involved in angiogenesis[17,18], such actions are considered minor or isolated. Therefore, we studied the effects of CCR3 modulation on angiogenesis *in vitro* and *in vivo*. Neutralizing anti-CCR3 antibodies inhibited the tube formation of primary human CECs cultured in Matrigel *in vitro* (Fig. 2a). In an experimental model of CNV induced by laser injury in wild-type mice[19–24], neutralizing anti-CCR3 antibodies reduced the fraction of CECs *in vivo* that was in the proliferative state of the cell cycle (Fig. 2b). Consistent with this finding, each of the three eotaxins stimulated human CEC proliferation (Fig. 2c). Cytoskeletal rearrangement through polymerization of monomeric actin to microfilamentous F-actin, which is essential for eosinophil chemotaxis induced by the eotaxins, is also critical in angiogenic migration of endothelial cells. Stimulation of human CECs with any of the three eotaxins induced a rapid polymerization of actin molecules (Fig. 2d, e). All three eotaxins also activated RAC1 (Supplementary Fig. 3), a small GTPase that is critical in regulating endothelial cell spreading and migration, and promoted human CEC migration in a dose-dependent fashion (Fig. 2f). Collectively, these data demonstrate that CCR3 activation can promote several steps of angiogenesis. The expression of CCR3 on CECs *in vivo* is confined to choroidal neovascular tissues; however, *in vitro*, human CECs responded to CCR3 ligands. This might be owing to the presence of several CNV-promoting growth factors in the culture medium.

CCR3 receptor or ligand antagonism inhibits CNV

We studied the *in vivo* effects of CCR3 targeting in a mouse model of CNV induced by laser injury[22], which is the most widely used animal model of this disease. A single intraocular administration of either CCR3-neutralizing antibodies or a small molecule CCR3 receptor antagonist ((*S*)-methyl-2-naphthoylamino-3-(4-nitrophenyl)propionate; SB328437) suppressed laser-injury-induced CNV in wild-type mice in a dose-dependent fashion (Fig. 3a–c). CNV was also diminished in $Ccr3^{-/-}$ mice[25] compared to wild-type mice (Fig. 3d). The specificity of pharmacological CCR3 blockade was confirmed by demonstrating that CNV was not reduced in $Ccr3^{-/-}$ mice by CCR3-neutralizing antibodies or a CCR3 receptor antagonist ($116 \pm 7\%$ and $109 \pm 16\%$ of control, respectively; $n = 5$; $P > 0.1$). CCL11 and CCL24, the principal mouse ligands for CCR3, were markedly increased soon after laser injury, and immunolocalized to the retinal pigmented epithelium (RPE), which is adjacent to CECs (Fig. 3e, f).

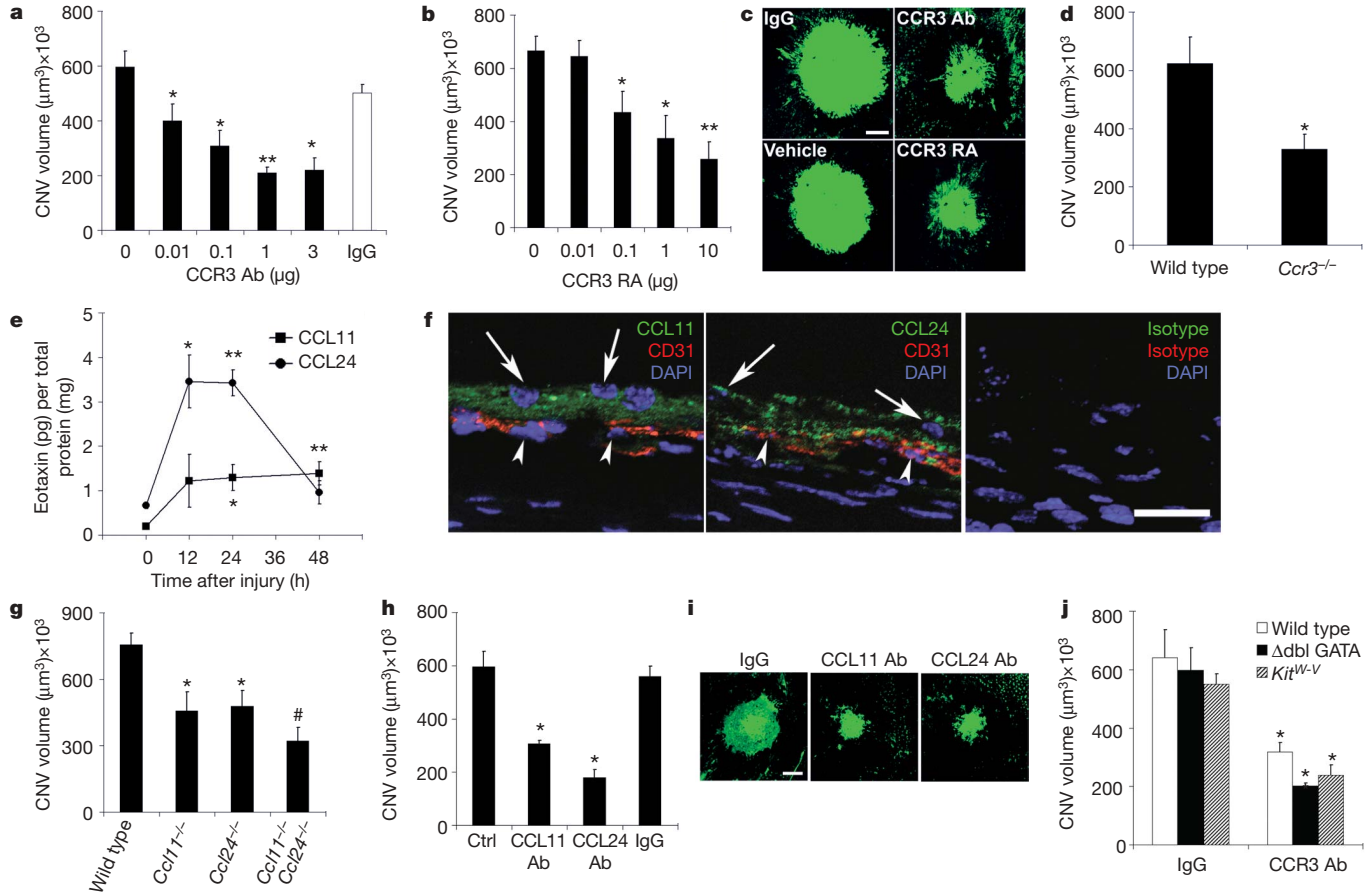


Figure 3 | CNV reduced by CCR3 or eotaxin ablation or blockade independent of leukocyte modulation. **a**, **b**, Laser-induced CNV in wild-type mice was reduced by neutralizing anti-mouse CCR3 antibody (CCR3 Ab) compared to isotype IgG (**a**) and by the CCR3 receptor antagonist (RA) SB328437 compared to vehicle (PBS/DMSO) (**b**) in a dose-dependent fashion. $n = 8$–12, $^*P < 0.05$, $^{**}P < 0.01$ compared to no antibody or receptor antagonist. **c**, Representative examples of CNV in drug-treated mice. **d**, Laser-induced CNV was reduced in $Ccr3^{-/-}$ mice compared to wild-type mice. $n = 9$, $^*P < 0.05$ compared to wild-type mice. **e**, Eotaxin-1 (CCL11) and eotaxin-2 (CCL24) protein levels, measured by ELISA, were increased after laser injury in wild-type mice. $n = 6$, $^*P < 0.05$ and $^{**}P < 0.01$ compared to the 0 h baseline. **f**, CCL11 and CCL24 immunofluorescence (green) was localized in the RPE cell layer (arrows) adjacent to $CD31^+$ (red) CECs (arrowheads) on day 1 after laser injury in wild-type mice. Nuclei were stained blue by DAPI. No specific immunofluorescence was detected with isotype control IgGs. Images representative of three independent experiments are shown. **g**, Laser-induced CNV was reduced in $Ccl11^{-/-}$ and $Ccl24^{-/-}$ mice compared to wild-type mice. $n = 8$–10, $^*P < 0.05$ compared to wild-type mice. CNV is further reduced in $Ccl11^{-/-}$ $Ccl24^{-/-}$ mice compared to single-null mice. $\#P < 0.05$ compared to single-null mice. **h**, Laser-induced CNV in wild-type mice was reduced by neutralizing antibodies against mouse CCL11 or CCL24 compared to isotype IgG. $n = 7$–10, $^*P < 0.05$ compared to no injection (control; Ctrl) or IgG. **i**, Representative examples of CNV in eotaxin-neutralized mice. **j**, Neutralizing anti-CCR3 antibodies reduced laser-induced CNV in mice deficient in eosinophils (ΔdblGATA) or mast cells ($Kit^{W\text{-}v}$). $n = 6$–9, $^*P < 0.05$ compared to IgG. Scale bars, 100 μm (**c**, **i**), and 20 μm (**f**). Error bars depict s.e.m.

Also, human RPE cells synthesized all three eotaxins (Supplementary Fig. 4), implicating these cells, which are abundantly interspersed in CNV[9], as a source of CCR3 ligands in CNV. Genetic ablation of either *Ccl11* (ref. 26) or *Ccl24* (ref. 12) reduced CNV, whereas the neovascular response in $Ccl11^{-/-}$ $Ccl24^{-/-}$ mice[12] was suppressed to a greater extent than in either of the single knockout mice, suggesting cooperation between these two ligands in this system (Fig. 3g). A single intraocular administration of neutralizing antibodies against CCL11 or CCL24 also suppressed CNV in wild-type mice (Fig. 3h, i), validating these CCR3 ligands as anti-angiogenic targets. Together, these data demonstrate that CCR3 activation is essential for *in vivo* angiogenesis in the most widely used preclinical model of neovascular AMD.

CCR3-driven angiogenesis uncoupled from inflammation

We sought to determine whether CCR3 targeting reduced CNV solely by anti-angiogenic mechanisms, or whether anti-inflammatory mechanisms were also involved. Neither eosinophils nor mast cells (defined as $CCR3^{hi}CD3^{-}CD117^{int}CD49d^{+}$ and $CCR3^{int}CD3^{-}CD117^{hi}CD49d^{+}$ cells, respectively) were recruited to the choroid after laser injury, as monitored by flow cytometry (Supplementary Fig. 5). Furthermore, the CNV response in eosinophil-deficient Δdbl GATA mice[11] (containing a deletion of the double GATA site) and mast-cell-deficient $Kit^{W\text{-}v}/Kit^{W\text{-}v}$ mice[27] was not different from the response in wild-type mice (Fig. 3j). Moreover, intraocular administration of neutralizing anti-CCR3 antibodies reduced CNV in Δdbl GATA or $Kit^{W\text{-}v}/Kit^{W\text{-}v}$ mice to the same extent as in wild-type mice. Thus, although eosinophils and mast cells have been reported to be capable of driving angiogenesis in other systems[17,18], both cell types are dispensable in the development of experimental CNV. Although neutrophil and macrophage infiltration are crucial for the development of experimental CNV[23,28], CCR3-receptor targeting did not affect recruitment of either inflammatory cell type (defined as $Gr\text{-}1^{+}F4/80^{-}$ and $F4/80^{+}CD11c^{-}$ cells, respectively; Supplementary Fig. 5). Therefore, the angioinhibitory effect of CCR3 blockade in this model is a direct anti-vascular effect, and does not seem to involve modulation of cellular inflammation. The mechanisms underlying the paucity of eosinophils and mast cells in CNV remain to be defined. One potential explanation could be the expression of CXCL9 in CNV, which blocks eotaxin-induced CCR3-mediated eosinophil recruitment (Supplementary Fig. 6)[29,30]. Other mechanisms influencing adhesion or mobilization of these leukocytes might also be operative.

CNV bioimaging by CCR3 targeting

Because invasion of the retina by CNV results in morphological and functional disruption of the retina, early detection of CNV is desirable; indeed, detection of CNV before retinal invasion would be ideal. CNV that has breached the retina can be detected by fluorescein angiography. However, this diagnostic modality cannot detect CNV before it has invaded the retina, that is, when it is still limited to the choroid. Yet, post-mortem histopathological studies have shown that substantial numbers of patients in whom fluorescein angiography does not reveal the presence of CNV nevertheless have CNV that has not yet invaded the retina[31]. Therefore, we explored whether CCR3-targeted bioimaging using anti-CCR3 Fab antibody fragments (Supplementary Fig. 7) conjugated to quantum dots (QDot-CCR3 Fab) could detect CNV before it became clinically evident.

We previously described the spontaneous development of CNV in senescent mice deficient in monocyte chemoattractant protein-1 (CCL2, also known as MCP-1) or its CCR2 receptor[32]. Similar pathology occurs at a younger age in $Ccl2^{-/-}$ $Ccr2^{-/-}$ mice (J.A., M.E.K., J.Z.B., H.K. and B.J.R., unpublished data). These mice also undergo outer retinal degeneration rapidly (Supplementary Fig. 8). We tested whether fundus angiography after intravenous injection of QDot-CCR3 Fab could detect subretinal CNV in these mice. QDot-CCR3 Fab angiography demonstrated hyperfluorescent signals in regions of the fundus of these mice that were silent on fluorescein angiography (Fig. 4a, b). The specificity of CCR3 targeting was confirmed by the absence of hyperfluorescent signals in $Ccl2^{-/-}$ $Ccr2^{-/-}$ mice injected with QDot-isotype Fab, and in wild-type mice injected with QDot-CCR3 Fab (Fig. 4b and Supplementary Fig. 9). Histological examination of these areas showed proliferating ($Ki67^{+}$) $CCR3^{+}$ blood vessels in the choroid that had not yet invaded the retina, along with the accumulation of QDot-CCR3 Fab in these vessels (Fig. 4c–g). These data provide proof-of-principle that CCR3-targeted bioimaging can detect subclinical CNV before it disrupts the retina and causes vision loss.

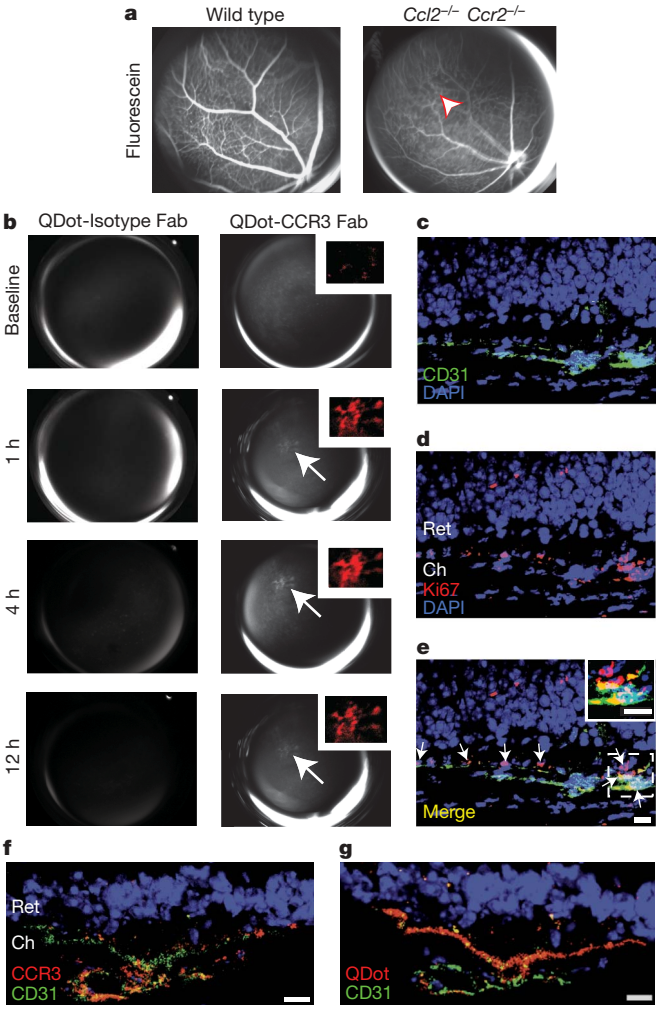


Figure 4 | CCR3-targeting quantum dots detect subretinal CNV. a, Images of the fundus taken after intravenous injection of sodium fluorescein in wild-type and $Ccl2^{-/-}$ $Ccr2^{-/-}$ mice show normal retinal vascular filling but no areas of hyperfluorescence indicative of CNV. **b**, After intravenous injection of QDot-CCR3 Fab in the same $Ccl2^{-/-}$ $Ccr2^{-/-}$ mouse shown in **a**, focal branching choroidal hyperfluorescence was visualized (arrow) at 1 h in the same area that was not hyperfluorescent during fluorescein angiography (arrowhead in **a**). The intensity of this hyperfluorescence (shown in red pseudocolour in the inset) increased, attaining a peak at 4 h, and then declined in intensity but still persisted at 12 h. Corresponding images of QDot-Isotype Fab angiography showed no hyperfluorescence. **c**–**e**, The region corresponding to the area of hyperfluorescence seen on QDot-CCR3 Fab angiography in **b** contained several $CD31^{+}$ blood vessels in the choroid (Ch) that were proliferating ($Ki67^{+}$; arrows) and had not invaded the retina (Ret). Individual green ($CD31^{+}$, **c**), red ($Ki67^{+}$, **d**), and merged (**e**) fluorescence channel images are shown. Nuclei were stained with DAPI (blue). Arrows point to proliferating endothelial cells. The inset in **e** shows $Ki67^{+}$ $CD31^{+}$ cells in higher magnification. **f**, QDot-CCR3 Fab hyperfluorescent areas were localized to areas of subretinal CNV with $CCR3^{+}$ endothelial cells. **g**, The QDot label was visualized within $CD31^{+}$ vasculature of subretinal CNV lesions. Images are representative of six independent experiments. Scale bars, 10 μm (**c**–**g**).

CCR3 targeting is superior to VEGF-A targeting

By comparing CCR3 targeting to VEGF-A targeting, the most effective approved treatment for human CNV, we found that CCR3-neutralizing antibodies were more effective than VEGF-A-neutralizing antibodies ($68 \pm 3\%$ versus $57 \pm 4\%$) at inhibiting laser-induced CNV in mice (Supplementary Fig. 10). In the laser-injury model, CCR3 neutralization did not change VEGF-A levels in the RPE/choroid and VEGF-A blockade did not change CCR3 expression on CECs (Supplementary Fig. 11): these two pathways seem to not be directly coupled. Repeated intravitreous administration of anti-VEGF-A antibodies resulted in anatomical and functional damage to the retina in wild-type mice (Supplementary Fig. 12), consistent with earlier reports that anti-VEGF-A therapy induces dysfunction in and damage to the inner and outer murine retina[6,7]. These effects were modest at a dose of anti-VEGF-A antibodies that suppressed mouse CNV, but more pronounced at a higher dose that is comparable to the dose used in humans. It should be noted that anti-VEGF-A pharmacotherapy has not been associated with an increased risk of profound retinal damage in humans[33], but subtle abnormalities have been observed[34,35] and some adverse effects might be misattributed to disease progression. In contrast to VEGF-A blockade, neither an anti-CCR3 antibody nor a CCR3 receptor antagonist induced retinal toxicity in wild-type mice, as confirmed by fundus imaging and electrophysiological function (Supplementary Fig. 12). *Vegfa* deletion is embryonically lethal[36,37] and conditional ablation of *Vegfa* in the RPE induces profound retinal degeneration and visual dysfunction[38]. In contrast, the $Ccr3^{-/-}$ mouse retina was normal in appearance and electrophysiological function (Supplementary Fig. 13).

Discussion

Our findings suggest that CCR3 targeting may be a safe and viable strategy for early detection (using biocompatible quantum dots or other bioimaging fluorochromes, such as near infrared dyes) and treatment of CNV (by receptor or ligand targeting), and might be superior to the current standard of care. CCR3 bioimaging is probably most useful in individuals with RPE pigmentary disturbances and multiple subretinal lipoproteinaceous deposits known as drusen or fellow eye involvement with clinically evident CNV, as they are known to be at high risk for developing CNV[39,40]. Similar techniques might be useful in non-invasively bioimaging other metabolic or molecular markers to provide information about disease pathogenesis or activity.

Several strategies have yielded molecular markers that are preferentially expressed on proliferating endothelial cells such as those in tumour vasculature[41,42]; however, CCR3 has not been identified in any of these reports. Therefore, our studies identify CCR3 as a new marker of pathological angiogenesis and as a functional target in neovascular AMD. These findings should also prompt a search for genetic polymorphisms in the eotaxin–CCR3 axis in patients with AMD, and investigations of CCR3 function in other models of angiogenesis. Also, it is tempting to speculate that targeting CCR3 might provide dual benefits in asthma, which involves varying degrees of eosinophilic inflammation as well as angiogenic airway remodelling[43].

METHODS SUMMARY

Mouse model of CNV. Laser photocoagulation (OcuLight GL, Iridex Corporation) was performed on mouse eyes to induce CNV, and CNV volumes were measured 7 days after injury by scanning laser confocal microscopy (TCS SP, Leica), as previously described[22].

Drug injections. Rat IgG2a neutralizing antibody against mouse CCR3 (R&D Systems), control rat IgG2a (Serotec), goat neutralizing antibody against mouse CCL11 (R&D Systems), goat neutralizing antibody against mouse CCL24 (R&D Systems), control goat IgG (Jackson Immunoresearch), or (*S*)-methyl-2-naphthoylamino-3-(4-nitrophenyl)propionate (SB328437; Calbiochem) dissolved in dimethylsulphoxide (DMSO) were injected into the vitreous humour using a 33-gauge double-calibre needle (Ito Corporation) once, immediately after laser injury as previously described[22].

CCR3 bioimaging. Fab fragments were created from monoclonal IgG2a antibody raised against the extracellular domain of murine CCR3 (R&D Systems) and an isotype rat IgG2a (R&D Systems) using a commercially available papain-based kit (Pierce). Recovered fragments were conjugated with quantum dots (Invitrogen, QDot-800) and resuspended in sterile PBS. $Ccl2^{-/-}$ $Ccr2^{-/-}$ mice were administered 100 μg of tagged CCR3 Fab or isotype Fab by tail-vein injection after acquiring baseline fluorescent imaging using a Topcon retinal camera (TRC-50IX). Serial images were then acquired at 1, 4 and 12 h, after which eyes were collected and frozen in OCT for immunofluorescent analyses. Retinal images were analysed (ImageNet, Topcon) by comparison to baseline and fluorescein angiographic data. Hyperfluorescent areas were then cropped, equally thresholded, and pseudocoloured (Photoshop CS3, Adobe). Sections from QDot-conjugated CCR3 or rat IgG2a isotype Fab injected animals were fixed in 4% paraformaldehyde and blocked with 5% normal donkey serum/5% goat serum in PBS, stained with rat anti-mouse CD31 (BD Biosciences) and either rabbit anti-mouse CCR3 (Santa Cruz) or rabbit anti-Ki67 (Abcam), followed by appropriate fluorescent secondary antibodies (Alexa Fluor 488/594, Invitrogen), and evaluated by confocal laser scanning microscopy (Leica SP-5).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 March; accepted 21 May 2009.
Published online 14 June 2009.

1. Ambati, J., Ambati, B. K., Yoo, S. H., Ianchulev, S. & Adamis, A. P. Age–related macular degeneration: etiology, pathogenesis, and therapeutic strategies. *Surv. Ophthalmol.* **48,** 257–293 (2003).
2. Gragoudas, E. S., Adamis, A. P., Cunningham, E. T. Jr, Feinsod, M. & Guyer, D. R. Pegaptanib for neovascular age-related macular degeneration. *N. Engl. J. Med.* **351,** 2805–2816 (2004).
3. Brown, D. M. *et al.* Ranibizumab versus verteporfin for neovascular age-related macular degeneration. *N. Engl. J. Med.* **355,** 1432–1444 (2006).
4. Rosenfeld, P. J. *et al.* Ranibizumab for neovascular age-related macular degeneration. *N. Engl. J. Med.* **355,** 1419–1431 (2006).
5. Famiglietti, E. V. *et al.* Immunocytochemical localization of vascular endothelial growth factor in neurons and glial cells of human retina. *Brain Res.* **969,** 195–204 (2003).
6. Nishijima, K. *et al.* Vascular endothelial growth factor-A is a survival factor for retinal neurons and a critical neuroprotectant during the adaptive response to ischemic injury. *Am. J. Pathol.* **171,** 53–67 (2007).
7. Saint-Geniez, M. *et al.* Endogenous VEGF is required for visual function: evidence for a survival role on muller cells and photoreceptors. *PLoS One* **3,** e3554 (2008).
8. Rothenberg, M. E. & Hogan, S. P. The eosinophil. *Annu. Rev. Immunol.* **24,** 147–174 (2006).
9. Submacular Surgery Trials Research Group. Histopathologic and ultrastructural features of surgically excised subfoveal choroidal neovascular lesions: submacular surgery trials report no. 7. *Arch. Ophthalmol.* **123,** 914–921 (2005).
10. Justice, J. P. *et al.* Ablation of eosinophils leads to a reduction of allergen-induced pulmonary pathology. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **284,** L169–L178 (2003).
11. Humbles, A. A. *et al.* A critical role for eosinophils in allergic airways remodeling. *Science* **305,** 1776–1779 (2004).
12. Pope, S. M., Zimmermann, N., Stringer, K. F., Karow, M. L. & Rothenberg, M. E. The eotaxin chemokines and CCR3 are fundamental regulators of allergen-induced pulmonary eosinophilia. *J. Immunol.* **175,** 5341–5350 (2005).
13. Jose, P. J. *et al.* Eotaxin: a potent eosinophil chemoattractant cytokine detected in a guinea pig model of allergic airways inflammation. *J. Exp. Med.* **179,** 881–887 (1994).
14. Teixeira, M. M. *et al.* Chemokine-induced eosinophil recruitment. Evidence of a role for endogenous eotaxin in an *in vivo* allergy model in mouse skin. *J. Clin. Invest.* **100,** 1657–1666 (1997).
15. Blanchard, C. *et al.* Eotaxin-3 and a uniquely conserved gene-expression profile in eosinophilic esophagitis. *J. Clin. Invest.* **116,** 536–547 (2006).
16. Salcedo, R. *et al.* Eotaxin (CCL11) induces *in vivo* angiogenic responses by human $CCR3^+$ endothelial cells. *J. Immunol.* **166,** 7571–7578 (2001).
17. Puxeddu, I. *et al.* Human peripheral blood eosinophils induce angiogenesis. *Int. J. Biochem. Cell Biol.* **37,** 628–636 (2005).
18. Heissig, B. *et al.* Low-dose irradiation promotes tissue revascularization through VEGF release from mast cells and MMP-9-mediated progenitor cell mobilization. *J. Exp. Med.* **202,** 739–750 (2005).
19. Tobe, T. *et al.* Targeted disruption of the FGF2 gene does not prevent choroidal neovascularization in a murine model. *Am. J. Pathol.* **153,** 1641–1646 (1998).
20. Nozaki, M. *et al.* Drusen complement components C3a and C5a promote choroidal neovascularization. *Proc. Natl Acad. Sci. USA* **103,** 2328–2333 (2006).
21. Nozaki, M. *et al.* Loss of SPARC-mediated VEGFR-1 suppression after injury reveals a novel antiangiogenic activity of VEGF-A. *J. Clin. Invest.* **116,** 422–429 (2006).
22. Kleinman, M. E. *et al.* Sequence- and target-independent angiogenesis suppression by siRNA via TLR3. *Nature* **452,** 591–597 (2008).

23. Sakurai, E., Anand, A., Ambati, B. K., van Rooijen, N. & Ambati, J. Macrophage depletion inhibits experimental choroidal neovascularization. *Invest. Ophthalmol. Vis. Sci.* **44,** 3578–3585 (2003).
24. Sakurai, E. *et al.* Targeted disruption of the CD18 or ICAM-1 gene inhibits choroidal neovascularization. *Invest. Ophthalmol. Vis. Sci.* **44,** 2743–2749 (2003).
25. Humbles, A. A. *et al.* The murine CCR3 receptor regulates both the role of eosinophils and mast cells in allergen-induced airway inflammation and hyperresponsiveness. *Proc. Natl Acad. Sci. USA* **99,** 1479–1484 (2002).
26. Rothenberg, M. E., MacLean, J. A., Pearlman, E., Luster, A. D. & Leder, P. Targeted disruption of the chemokine eotaxin partially reduces antigen-induced tissue eosinophilia. *J. Exp. Med.* **185,** 785–790 (1997).
27. Kitamura, Y., Go, S. & Hatanaka, K. Decrease of mast cells in W/W^v mice and their increase by bone marrow transplantation. *Blood* **52,** 447–452 (1978).
28. Zhou, J. *et al.* Neutrophils promote experimental choroidal neovascularization. *Mol. Vis.* **11,** 414–424 (2005).
29. Fulkerson, P. C. *et al.* Negative regulation of eosinophil recruitment to the lung by the chemokine monokine induced by IFN-γ (Mig, CXCL9). *Proc. Natl Acad. Sci. USA* **101,** 1987–1992 (2004).
30. Fulkerson, P. C., Zhu, H., Williams, D. A., Zimmermann, N. & Rothenberg, M. E. CXCL9 inhibits eosinophil responses by a CCR3- and Rac2-dependent mechanism. *Blood* **106,** 436–443 (2005).
31. Green, W. R. & Key, S. N. III. Senile macular degeneration: a histopathologic study. *Trans. Am. Ophthalmol. Soc.* **75,** 180–254 (1977).
32. Ambati, J. *et al.* An animal model of age-related macular degeneration in senescent Ccl-2- or Ccr-2-deficient mice. *Nature Med.* **9,** 1390–1397 (2003).
33. Ip, M. S. *et al.* Anti-vascular endothelial growth factor pharmacotherapy for age-related macular degeneration: a report by the American Academy of Ophthalmology. *Ophthalmology* **115,** 1837–1846 (2008).
34. Sayanagi, K., Sharma, S. & Kaiser, P. K. Photoreceptor status after anti-vascular endothelial growth factor therapy in exudative age-related macular degeneration. *Br. J. Ophthalmol.* **93,** 622–626 (2009).
35. Yodoi, Y. *et al.* Central retinal sensitivity after intravitreal injection of bevacizumab for myopic choroidal neovascularization. *Am. J. Ophthalmol.* **147,** 816–824 (2009).
36. Carmeliet, P. *et al.* Abnormal blood vessel development and lethality in embryos lacking a single VEGF allele. *Nature* **380,** 435–439 (1996).
37. Ferrara, N. *et al.* Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene. *Nature* **380,** 439–442 (1996).
38. Marneros, A. G. *et al.* Vascular endothelial growth factor expression in the retinal pigment epithelium is essential for choriocapillaris development and visual function. *Am. J. Pathol.* **167,** 1451–1459 (2005).
39. Bressler, S. B., Maguire, M. G., Bressler, N. M. & Fine, S. L. (The Macular Photocoagulation Study Group). Relationship of drusen and abnormalities of the retinal pigment epithelium to the prognosis of neovascular macular degeneration. *Arch. Ophthalmol.* **108,** 1442–1447 (1990).
40. Macular Photocoagulation Study Group. Risk factors for choroidal neovascularization in the second eye of patients with juxtafoveal or subfoveal choroidal neovascularization secondary to age-related macular degeneration. *Arch. Ophthalmol.* **115,** 741–747 (1997).
41. St Croix, B. *et al.* Genes expressed in human tumor endothelium. *Science* **289,** 1197–1202 (2000).
42. Zhang, L. *et al.* Gene expression profiles in normal and cancer cells. *Science* **276,** 1268–1272 (1997).
43. Wenzel, S. E. Eosinophils in asthma—closing the loop or opening the door? *N. Engl. J. Med.* **360,** 1026–1028 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. King, L. Xu, M. McConnell, K. Emerson, G. R. Pattison and M. Mingler for technical assistance, J. M. Farber for the gift of a reagent, R. J. Kryscio for statistical guidance, and B. Appukuttan, M. W. Fannon, R. Mohan, A. P. Pearson, A. M. Rao, G. S. Rao and K. Ambati for discussions. J.A. was supported by National Eye Institute/National Institutes of Health (NIH) grants EY015422, EY018350 and EY018836, the Doris Duke Distinguished Clinical Scientist Award, the Burroughs Wellcome Fund Clinical Scientist Award in Translational Research, the Macula Vision Research Foundation, the E. Matilda Ziegler Foundation for the Blind, the Dr. E. Vernon Smith and Eloise C. Smith Macular Degeneration Endowed Chair, the Lew R. Wassermann Merit & Physician Scientist Awards (Research to Prevent Blindness, RPB), the American Health Assistance Foundation, and a departmental unrestricted grant from the RPB. J.Z.B. was supported by the University of Kentucky Physician Scientist Award. M.E.K. was supported by the International Retinal Research Foundation Dr. Charles Kelman Postdoctoral Scholar Award. R.J.C.A. was supported by Fight for Sight. B.K.A. was supported by NIH grants EY017182 and EY017950, the VA Merit Award and the Department of Defense. M.E.R. was supported by NIH grants AI45898 and DK076893. C.J.G. was supported by NIH grant AI039759. M.E.H. was supported by NIH grants EY017011 and EY015130 and a RPB departmental unrestricted grant. J.R.S. was supported by NIH grant EY010572, and RPB Career Development Award and a departmental unrestricted grant.

Author Contributions A.T., J.Z.B., M.E.K., W.G.C., M.N., K.Y., H.K., R.J.C.A., S.D., K.S., B.J.R., M.G.G., S.J.B., P.G. and A.M. performed experiments. S.G., A.A.H., Y.P., J.D.W., J.R.S., Y.O. and T.I. provided reagents. J.A. conceived and directed the project, and, with assistance from B.K.A., M.E.H., M.E.R., R.J.C.A. and J.R.S., wrote the paper. All authors had the opportunity to discuss the results and comment on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.A. (jamba2@email.uky.edu).

METHODS

Human tissue. Choroidal neovascular tissue was excised from patients with AMD who had no prior treatment for CNV. Retinal fibrosis tissue was excised from patients with a diagnosis of epiretinal membrane formation. Donor eyes from patients with atrophic AMD without CNV and patients without AMD were obtained from eye banks. Eyes with choroidal melanoma were obtained by surgical enucleation. The study followed the guidelines of the Declaration of Helsinki. Institutional review boards granted approval for allocation and histological analysis of specimens.

Animals. All animal experiments were in accordance with the guidelines of the University of Kentucky Institutional Animal Care and Use Committee and the Association for Research in Vision and Ophthalmology. C57BL/6J and $Kit^{W\text{-}v}$/$Kit^{W\text{-}v}$ mice were purchased from The Jackson Laboratory. $Ccr3^{-/-}$, $Ccl11^{-/-}$, $Ccl24^{-/-}$, $Ccl11^{-/-}$ $Ccl24^{-/-}$ and Δdbl GATA mice have been previously described[11,12,25,26]. $Ccl2^{-/-}$ $Ccr2^{-/-}$ mice were generated by interbreeding single knockout mice described previously[32].

Drug injections. Rat IgG2a neutralizing antibody against mouse CCR3 (R&D Systems), control rat IgG2a (Serotec), goat neutralizing antibody against mouse CCL11 (1 μg; R&D Systems), goat neutralizing antibody against mouse CCL24 (5 μg; R&D Systems), control goat IgG (Jackson Immunoresearch), or (*S*)-methyl-2-naphthoylamino-3-(4-nitrophenyl)propionate (SB328437; Calbiochem) dissolved in DMSO were injected into the vitreous humour of mice using a 33-gauge double-calibre needle (Ito Corporation) once, immediately after laser injury as previously described[22].

Flow cytometry. Rat antibody against mouse CCR3 (1:250; Santa Cruz) coupled with phycoerythrin (PE)-donkey antibody against rat IgG (1:250; Jackson Immunoresearch) or AlexaFluor647-conjugated rat antibody against mouse CCR3 (10 μg ml^{-1}; BD Biosciences) were used to quantify cell surface receptor expression on CECs, defined by CD31$^+$ VEGFR-2$^+$ expression, gated by FITC-conjugated rat antibody against mouse CD31 (20 μg ml^{-1}; BD Biosciences) and PE-conjugated rat antibody against mouse VEGFR-2 (20 μg ml^{-1}; BD Biosciences). Macrophages, neutrophils, eosinophils and mast cells were defined as F4/80$^+$ CD11c$^-$, Gr-1$^+$ F4/80$^-$, CCR3^{hi} CD3$^-$ CD117^{int} CD49d$^+$ and CCR3^{int} CD3$^-$ CD117^{hi} CD49d$^+$ cells, respectively. The DNA content for cell cycle was analysed after incubation with propidium iodide (0.05 mg ml^{-1}; Molecular Probes) containing 0.1% Triton X-100 and RNase A (0.1 mg ml^{-1}; Roche). Samples were analysed on a LSRII (Becton Dickinson).

Immunolabelling. Immunofluorescent staining was performed with antibodies against human CCR3 (rat monoclonal, R&D Systems) or human CD31 (mouse monoclonal, Dako), and identified with Alexa 488 (Molecular Probes) or Cy3 secondary antibodies (Jackson ImmunoResearch). Immunohistochemical staining with the primary antibodies specific for human eotaxins-1, -2 and -3 (mouse monoclonal, R&D Systems) was performed using horseradish peroxidase. Laser-injured mouse eye sections were stained with antibodies against mouse CCL11 or CCL24 (both R&D Systems) along with antibody against mouse CD31 (BD Biosciences) and visualized with FITC or Cy3 secondary antibodies. Images were obtained using Leica SP5 or Zeiss Axio Observer Z1 microscopes.

Tube formation assay. Ninety-six-well plates were coated with Growth-Factor-Reduced Matrigel (BD Biosciences) mixed with rat neutralizing-antibody against human CCR3 (20 μg ml^{-1}, R&D Systems) or control rat IgG2a (Invitrogen) and allowed to solidify in the incubator at 37 °C for 45 min. Human CECs[44–47] were plated on top of the Matrigel at 2.25×10^4 cm^{-2} in EBM-2 basal media (Cambrex) containing 1% FBS with CCR3 antibody or rat IgG2a at the concentrations shown and allowed to grow overnight. Tube formation was analysed by counting the number of cell junctions per mm^2.

Proliferation assay. Human CECs were synchronized for cell cycle state by first cultivating them in EGM2-MV media (Lonza) supplemented with 10% FBS (Gibco) to achieve complete confluence, and then by overnight serum starvation in MCDB131 media (Gibco) with 0.1% FBS. They were passaged to 96-well plates at a density of 5,000 cells per well, followed by stimulation for 24 h with eotaxin-1, 2 or 3 (10 ng, 100 ng and 2 μg per ml, respectively; Peprotech) in MCDB131 media with 0.1% FBS. After 24 h, cell viability was measured with BrdU ELISA (Chemicon) according to the manufacturer's instructions.

F-actin polymerization assay. Human CECs were seeded in black-walled 96-well plates and grown to 70–80% confluence in fully supplemented EGM-2MV. Cultures were serum-starved overnight in basal media and then stimulated with recombinant human eotaxin-1 (10 ng ml^{-1}), eotaxin-2 (100 ng ml^{-1}), eotaxin-3 (2 μg ml^{-1}) (Peprotech), or vehicle control (PBS). At 0, 10, 30, 60 or 120 s timepoints, cells were fixed in 3.7% paraformaldehyde for 10 min, washed, permeabilized in PBS with 0.1% Triton X-100, and then stained with rhodamine-labelled phalloidin (1:200, Invitrogen) as per the manufacturer's recommendations. Plates were analysed on a fluorescent plate reader (Synergy 4, Biotek) followed by fluorescent microscopy (Nikon E800).

Migration assay. Eotaxins-1, -2 and -3 were reconstituted in 0.1% BSA and then mixed with Matrigel diluted 1:1 with serum-free endothelial basal media (EBM-2; Lanza). Five-hundred microlitres of EBM-2 was added to each well of a 24-well plate, followed by a 6.5-mm diameter Transwell insert (8 μm pores; Corning). Human CECs in EBM-2 were prestained with Vybrant DiO (Invitrogen) for 30 min at 37 °C and seeded into the inserts at 50,000 cells per 200 μl of serum free EBM-2 media. The plates were allowed to incubate for 16 h at 37 °C, 5% CO_2. The migrated cells were imaged with an Olympus CK40 microscope and Olympus DP71 camera.

RAC1 activation. Human CECs were cultured in EGM-2 MV containing 5% FBS. Before starting the assay, cells were serum-starved overnight using basal medium (MCDB131) supplemented with 1% FBS. Cells were stimulated for designated times with eotaxin-1, -2 and -3 (10 ng ml^{-1}, 100 ng ml^{-1} and 2 μg ml^{-1}, respectively). Equal amounts of lysates (500 μg) were incubated with GST–Pak1–PBD agarose beads (Upstate) to pull down active GTP-bound RAC1 at 4 °C for 1 h with rotation. The samples were subsequently analysed for bound RAC1 by western blot analysis using an anti-RAC1 antibody (Upstate).

Electroretinography. Mice were dark-adapted overnight and then anaesthetized. Both eyes were positioned within a ColourBurst Ganzfeld stimulator (Diagnosys). Espion software (Diagnosys) was used to program a fully automated flash intensity series, from which retinal responses were recorded.

44. Geisen, P., McColm, J. R. & Hartnett, M. E. Choroidal endothelial cells transmigrate across the retinal pigment epithelium but do not proliferate in response to soluble vascular endothelial growth factor. *Exp. Eye Res.* **82**, 608–619 (2006).
45. Peterson, L. J., Wittchen, E. S., Geisen, P., Burridge, K. & Hartnett, M. E. Heterotypic RPE-choroidal endothelial cell contact increases choroidal endothelial cell transmigration via PI 3-kinase and Rac1. *Exp. Eye Res.* **84**, 737–744 (2007).
46. Smith, J. R. *et al.* Unique gene expression profiles of donor-matched human retinal and choroidal vascular endothelial cells. *Invest. Ophthalmol. Vis. Sci.* **48**, 2676–2684 (2007).
47. Zamora, D. O. *et al.* Proteomic profiling of human retinal and choroidal endothelial cells reveals molecular heterogeneity related to tissue of origin. *Mol. Vis.* **13**, 2058–2065 (2007).

ARTICLES

Mechanisms promoting translocations in editing and switching peripheral B cells

Jing H. Wang[1,2,3,4]*, Monica Gostissa[1,2,3,4]*, Catherine T. Yan[1,2,3,4]*, Peter Goff[1,2,3,4], Thomas Hickernell[1,2,3,4], Erica Hansen[1,2,3,4], Simone Difilippantonio[5], Duane R. Wesemann[1,2,3,4,6], Ali A. Zarrin[1,2,3,4]†, Klaus Rajewsky[3], Andre Nussenzweig[5] & Frederick W. Alt[1,2,3,4]

Variable, diversity and joining gene segment (V(D)J) recombination assembles immunoglobulin heavy or light chain (IgH or IgL) variable region exons in developing bone marrow B cells, whereas class switch recombination (CSR) exchanges IgH constant region exons in peripheral B cells. Both processes use directed DNA double-strand breaks (DSBs) repaired by non-homologous end-joining (NHEJ). Errors in either V(D)J recombination or CSR can initiate chromosomal translocations, including oncogenic IgH locus (*Igh*) to *c-myc* (also known as *Myc*) translocations of peripheral B cell lymphomas. Collaboration between these processes has also been proposed to initiate translocations. However, the occurrence of V(D)J recombination in peripheral B cells is controversial. Here we show that activated NHEJ-deficient splenic B cells accumulate V(D)J-recombination-associated breaks at the lambda IgL locus (*Igl*), as well as CSR-associated *Igh* breaks, often in the same cell. Moreover, *Igl* and *Igh* breaks are frequently joined to form translocations, a phenomenon associated with specific *Igh–Igl* co-localization. *Igh* and *c-myc* also co-localize in these cells; correspondingly, the introduction of frequent *c-myc* DSBs robustly promotes *Igh–c-myc* translocations. Our studies show peripheral B cells that attempt secondary V(D)J recombination, and determine a role for mechanistic factors in promoting recurrent translocations in tumours.

Recombination activating gene 1/2 (RAG1/2) endonucleases initiate V(D)J recombination by cleaving V, D and J segments, which are joined exclusively by NHEJ to form V(D)J exons[1,2]. V(D)J recombination in bone marrow pro-B cells first assembles IgH V(D)J exons leading to μ chain expression[3]. Subsequently, IgL VJ exons are assembled in pre-B cells, generating immature B cells that express μ plus IgL chains as surface IgM[3]. The two types of IgL proteins (Igκ and Igλ) are encoded in distinct loci (*Igk* and *Igl*), and primary *Igk* V(D)J recombination usually precedes that of *Igl*[4]. Individual B cells express either Igκ or Igλ, with about 95% of mouse IgM^+ B cells being $Ig\kappa^+$ and the remainder $Ig\lambda^+$ (ref. 4). Newly generated bone marrow B cells that express autoreactive receptors can undergo tolerogenic secondary V(D)J recombination, termed receptor editing, in which they further rearrange or delete *Igk* and may rearrange *Igl*[5–7] (see Supplementary Fig. 1 for a schematic of these processes).

Surface IgM^+ B cells downregulate RAG and migrate to peripheral lymphoid tissues (for example, spleen) where they participate in antigen-dependent responses including CSR[8]. The various sets of germline *Igh* constant region exons ('C_H genes') are flanked by switch (S) regions[9]. Activation-induced cytidine deaminase (AID, also known as AICDA) initiates DSBs in both Sμ and a downstream S region, which are then joined by NHEJ or, in its absence, by less efficient microhomology-mediated alternative end-joining[9,10]. This replaces Cμ with a downstream C_H gene to complete CSR (Supplementary Fig. 1). Germinal centre B cells have been argued to undergo antigen-dependent secondary V(D)J recombination, termed receptor revision, as a means of diversification[11]. Like receptor editing, receptor revision is proposed to target *Igk* and *Igl*, but to be distinct in location and activation mechanism[11,12]. Whether or not V(D)J recombination actually occurs in the context of receptor revision in germinal centre B cells has been debated[11–14].

Human and mouse B lymphomas often contain clonal translocations linking oncogenes, such as *c-myc*, to *Igh*, *Igk* or *Igl*[15,16]. Such recurrent oncogenic translocations are thought to represent highly selected, very low frequency events. Even so, aspects of *c-myc*, beyond coding sequences, may increase its translocation fequency[17]. In this regard, loci involved in recurrent oncogenic translocations often are spatially proximal within interphase nuclei[18–23]. RAG and AID have been implicated in collaboratively initiating oncogenic translocations in human bone-marrow-derived pro-B/pre-B lymphomas[24,25]. Many oncogenic translocations in mature B lymphomas occur during attempted CSR and involve AID-initiated breaks[26–29]; but others result from RAG-initiated DSBs[15,30,31]. Owing to checkpoint defects, RAG-initiated *Igh* breaks in ATM-deficient bone marrow pro-B cells persist and can be translocation substrates in IgM^+ peripheral B cells[32]. Thus far, however, translocations have not been shown to result from RAG activity in peripheral B cells.

XRCC4 is a critical NHEJ component[2]. In its absence, V(D)J recombination is abrogated[33,34] and CSR is impaired[10,35]. Conditional inactivation of *loxP*-flanked *Xrcc4* in p53-deficient peripheral B cells using a CD21-Cre transgene leads to recurrent CD21-Cre-inactivated XRCC4 plus p53-deficient ('CXP') B cell lymphomas that contain aberrant *Igk* and *Igl* V(D)J rearrangements, *Igh* CSR events, and *Igl* and/or *Igh–c-myc* translocations[36]. We proposed CXP tumour progenitors to be peripheral B cells that undergo secondary V(D)J recombination and CSR[36]. To search for such putative CXP tumour progenitors, we have now analysed splenic *CD21-cre*, $Xrcc4^{c/-}$ B cells, in which *Xrcc4* was peripherally inactivated but *p53* was left intact to obviate B cell lymphomas.

[1]Howard Hughes Medical Institute, [2]The Children's Hospital, [3]Immune Disease Institute, [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Experimental Immunology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [6]Division of Rheumatology, Allergy and Immunology, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. †Present address: Ali A. Zarrin, Immunology Discovery Group, Genentech, South San Francisco, California 94080, USA.
*These authors contributed equally to this work.

Igh breaks in XRCC4-deficient B cells are AID-dependent

CD21-cre, *Xrcc4*$^{c/-}$ mice have normal IgM$^+$ B cell numbers, because *Xrcc4* is intact for primary V(D)J recombination in developing bone marrow B cells, with inactivation starting in transitional-stage peripheral B cells[10,37]. *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells activated for CSR have high levels of *Igh* breaks on chromosome 12 owing to impaired NHEJ[10]. Although XRCC4 deficiency is not associated with known checkpoint defects[34], we tested the AID-dependency of *CD21-cre*, *Xrcc4*$^{c/-}$ B cell *Igh* breaks by breeding this genotype onto an AID-deficient (*Aid*$^{-/-}$) background[38] to generate *CD21-cre*, *Xrcc4*$^{c/-}$ *Aid*$^{-/-}$ mice. We stimulated *CD21-cre*, *Xrcc4*$^{c/-}$, *CD21-cre*, *Xrcc4*$^{c/-}$ *Aid*$^{-/-}$ and control (*Xrcc4*$^{c/c}$) splenic B cells with anti-CD40 plus IL4 for 4 days to promote IgG1 CSR, and assayed metaphases for *Igh* breaks and translocations by fluorescence *in situ* hybridization (FISH) with 5′ and 3′ *Igh* probes. Although general chromosomal breaks, as expected, were largely AID-independent in activated *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells (Supplementary Table 2), most of the *Igh* breaks were AID-dependent (Fig. 1a and Supplementary Table 1).

RAG-dependent *Igl* breaks in XRCC4-deficient B cells

We assayed activated *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells for *Igl* breaks using metaphase FISH with 5′ and 3′ *Igl* probes that flank the 200-kb *Igl* locus on chromosome 16 (Fig. 1b). After stimulation with anti-CD40 and IL4 for 4 days, we found *Igl* breaks in more than 1% of XRCC4-deficient B cells, with none in controls (Fig. 1b and Supplementary Table 3). Moreover, the *Igl* breaks were frequently translocated (Fig. 1b and Supplementary Fig. 2). Metaphase FISH with BAC probes flanking *Igk* showed that 1% of activated *CD21-cre*, *Xrcc4*$^{c/-}$ B cells also contain *Igk* breaks/translocations (Fig. 1c, Supplementary Fig. 2 and Supplementary Table 4). In contrast, XRCC4-deficient embryonic stem (ES) cells lacked *Igk* or *Igl* abnormalities (Supplementary Table 5). To determine when *Igl* and *Igk* breaks occurred, we assayed *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells at days 2, 3 and 4 of activation and found that both accumulated during stimulation, with *Igk* breaks kinetically preceding *Igl* breaks (Supplementary Tables 3, 4 and 6 and Supplementary Fig. 3). We also assayed for *Igl* breaks by three-dimensional (3D) interphase FISH with 5′ *Igl* and 3′ *Igl* probes (Fig. 1d). *Igl* breaks were rare in resting (day 0) *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells, but occurred in about 1.5% of day-4-activated XRCC4-deficient splenic B cells (Fig. 1d, Supplementary Fig. 4 and Supplementary Table 7). Thus, *Igl* and *Igk* breaks occur during expansion of activated *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells, a conclusion supported by our findings that p53 deficiency did not markedly enhance *Igl* breaks (Supplementary Table 3), and that 50% of metaphases with *Igl* breaks retained the acentric chromosome 16 fragment (Fig. 1b and data not shown).

To test AID involvement, we assayed for *Igk* and *Igl* breaks in day 4 *CD21-cre*, *Xrcc4*$^{c/-}$ *Aid*$^{-/-}$ B cells activated with anti-CD40 and IL4,

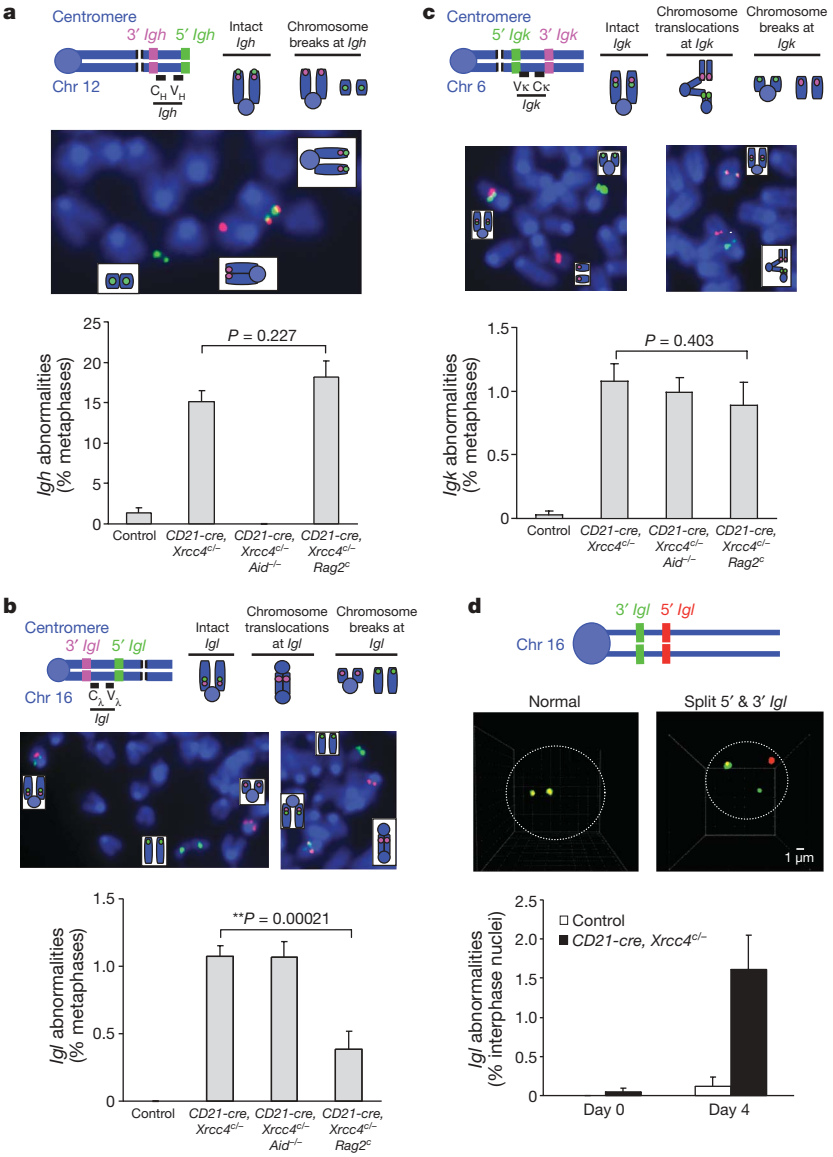


Figure 1 | Role of AID and RAG in generating *Igh*, *Igk* and *Igl* breaks in *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells. **a**, Top, diagram of *Igh* FISH probes. An intact *Igh* shows co-localized red and green signals; a broken locus appears as split red and green signals. Chr, chromosome. Middle, metaphase FISH showing *Igh* breaks. Bottom, quantification of *Igh* abnormalities in day 4 control ($n = 6$), *CD21-cre*, *Xrcc4*$^{c/-}$ ($n = 9$), *CD21-cre*, *Xrcc4*$^{c/-}$ *Aid*$^{-/-}$ ($n = 5$) and *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*^{c} ($n = 8$) splenic B cells activated with anti-CD40 and IL4 (see Supplementary Table 1). **b**, Top, diagram of *Igl* FISH probes. Green and red signals indicate co-localization as described in **a**. Middle, metaphase FISH showing *Igl* breaks (left) and an *Igl* break and dicentric translocation (right). Bottom, quantification of *Igl* abnormalities in day 4 control ($n = 11$), *CD21-cre*, *Xrcc4*$^{c/-}$ ($n = 11$), *CD21-cre*, *Xrcc4*$^{c/-}$ *Aid*$^{-/-}$ ($n = 3$) and *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*^{c} ($n = 8$) splenic B cells stimulated with anti-CD40 and IL4 (see Supplementary Table 3). **c**, Top, diagram of *Igk* FISH probes. *Igk* breaks are scored similarly as *Igl* breaks. Middle, metaphase FISH showing an *Igk* break (left) and *Igk* break and translocations (right), involving both centromeric and telomeric portions of chromosome 6. Bottom, quantification of *Igk* abnormalities in day 4 anti-CD40- and IL4-activated control ($n = 10$), *CD21-cre*, *Xrcc4*$^{c/-}$ ($n = 11$), *CD21-cre*, *Xrcc4*$^{c/-}$ *Aid*$^{-/-}$ ($n = 3$) and *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*^{c} ($n = 7$) splenic B cells (see Supplementary Table 4). **d**, Top, diagram of *Igl* 3D interphase FISH probes. Middle, 3D interphase FISH showing intact (co-localization of green and red signals) and broken (split green and red signals) *Igl* (see Supplementary Fig. 4). Bottom, quantification of *Igl* abnormalities by 3D interphase FISH on day 0 ($n = 3$) or day 4 ($n = 3$) splenic B cells activated with anti-CD40 and IL4. In all panels, data are presented as mean and s.e.m. Statistical analyses were calculated by a Student's *t*-test with two-tailed distribution.

and found a comparable frequency to that in *CD21-cre*, *Xrcc4*$^{c/-}$ B cells (Fig. 1b, c, Supplementary Tables 3 and 4 and Supplementary Fig. 5). Similar to earlier studies[32,39], we only found very low RAG expression in activated normal and *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells (data not shown). To further assess RAG involvement, we bred a *loxP*-flanked *Rag2* conditional allele (*Rag2*$^{c/c}$)[40] into the *CD21-cre*, *Xrcc4*$^{c/-}$ genotype to generate *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*$^{c/c}$ or *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*$^{c/-}$ ('*CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*^{c}') mice. After activation, the *Rag2* conditional allele was largely deleted in days 3- and 4-activated *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*^{c} cells (Supplementary Fig. 6). Although *Igh* break frequency was comparable between *CD21-cre*, *Xrcc4*$^{c/-}$ and *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*^{c} B cells, *Igl* break frequency was significantly reduced in *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*^{c} B cells (Fig. 1a, b and Supplementary Tables 1 and 3). Thus, in activated *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells, *Igh* breaks are AID-dependent and RAG-independent, whereas *Igl* breaks are AID-independent and mostly RAG-dependent. *Igk* breaks were not significantly reduced in activated *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*^{c} B cells (Fig. 1c and Supplementary Table 4), suggesting that either they are not initiated by AID or RAG, or their earlier kinetic onset allows accumulation before RAG activity is eliminated.

RAG and AID collaborate to generate *Igh–Igl* translocations

We used sequential FISH to determine whether *Igh*, *Igk* or *Igl* breaks occurred simultaneously in *CD21-cre*, *Xrcc4*$^{c/-}$ B cell metaphases. Analyses of over 2,000 day 4 *CD21-cre*, *Xrcc4*$^{c/-}$ B cell metaphases activated with anti-CD40 and IL4 showed that none had both *Igk* and *Igl* breaks (Supplementary Fig. 7). However, analyses with a Jκ–Cκ probe showed that nearly 50% of metaphases with a broken *Igl* had deleted Jκ–Cκ on one or both alleles (Supplementary Fig. 8), similar to secondary V(D)J recombination events in CXP B lymphomas[36]. We found one *Igk–Igh* translocation in over 2,000 activated *CD21-cre*, *Xrcc4*$^{c/-}$ B cell metaphases, consistent with a high frequency but at levels just below ready cytogenetic measurement (Supplementary Fig. 9a). Nearly 60% of metaphases with *Igl* breaks also had *Igh* breaks and/or translocations and about 20% of these retained both centric and acentric portions of chromosomes 12 and 16 (Supplementary Figs 7, 9b, c and data not shown), suggesting attempted V(D)J recombination and CSR in the same or successive cell cycles. In this regard, combined FISH with *Igl* and *Igh* probes and chromosome paints showed that 30% of *Igl* translocations involved *Igh* (for example, Fig. 2a and Supplementary Fig. 9b, c). Because many *Igh–Igl* translocations resulted in dicentrics with 3′ *Igl* and 3′ *Igh* probes juxtaposed (Supplementary Fig. 9b), we performed FISH with these probes simultaneously, which showed AID-dependent *Igh–Igl* translocations in about 0.2% of *CD21-cre*, *Xrcc4*$^{c/-}$ B cells (Fig. 2a, Supplementary Table 8 and Supplementary Fig. 10). Thus, unrepaired RAG-dependent *Igl* breaks in activated *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells are frequently fused to AID-dependent *Igh* breaks in the same cell to form chromosome 12/16 translocations.

We isolated *Igh–Igl* translocation junctions from *CD21-cre*, *Xrcc4*$^{c/-}$ B cells using PCR (Supplementary Fig. 10), and found most fused Sμ to sequences downstream of Jλ1/Jλ3 V(D)J recombination signal sequences (Fig. 2b and Supplementary Fig. 11). Consistent with AID-initiated *Igh* breaks joined to RAG-initiated *Igl* breaks, point mutations and other alterations were observed in *Igh*- but not *Igl*-derived junctional sequences (Supplementary Fig. 11). Consistent with RAG-initiated breaks resolved in the absence of NHEJ, *Igl* junctions were at variable distances downstream of Jλ1 and Jλ3. Furthermore, most *Igh/Igl* junctions contained microhomologies indicative of alternative end-joining (Supplementary Fig. 11). We conclude that, in activated splenic *CD21-cre*, *Xrcc4*$^{c/-}$ B cells, alternative end-joining joins RAG-induced *Igl* breaks to AID-initiated *Igh* breaks at high frequency.

Co-localization of *Igh* and *Igl* in B cell interphase nuclei

3D interphase FISH with 3′ *Igh* and 3′ *Igl* probes showed co-localization of the loci (≤0.5 μm apart) in about 14% of resting (day 0) and 7–8% of day 3.5 anti-CD40- and IL4-activated control and *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells (Fig. 3a–d and Supplementary Tables 9 and 10). Because there are no *Igh* or *Igl* breaks in resting B cells (Fig. 1a, d), and AID-initiated breaks begin at day 2 (ref. 41), we

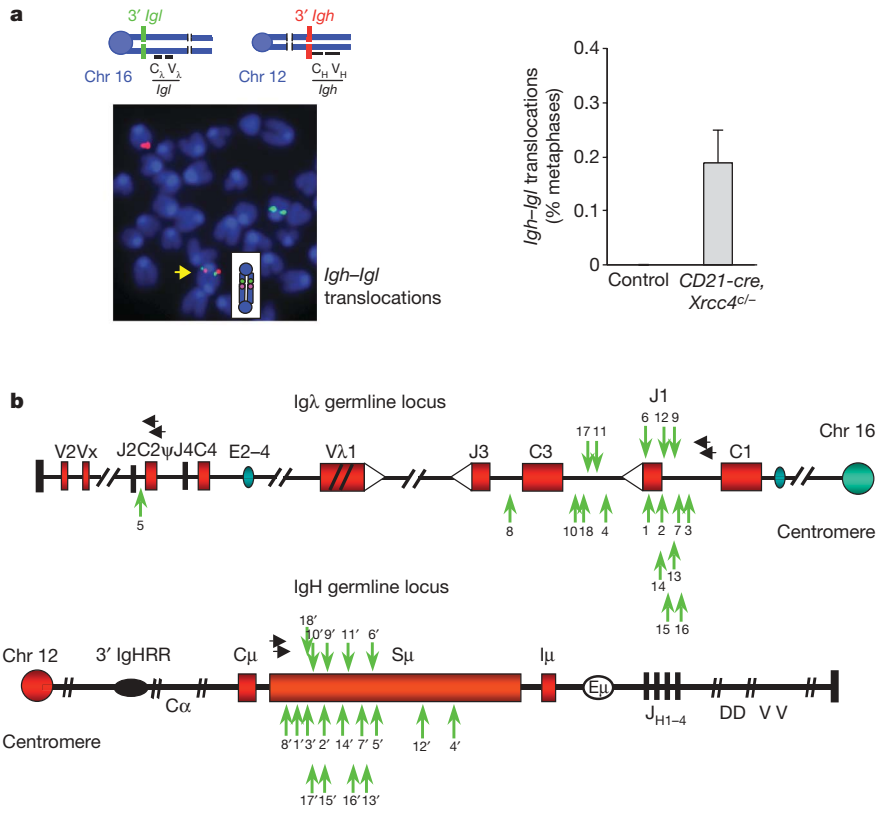


Figure 2 | Frequent *Igh–Igl* translocations in activated XRCC4-deficient splenic B cells. a, Top left, diagram showing 3′ *Igl* probe (green) on chromosome 16 and 3′ *Igh* probe (red) on chromosome 12. Bottom left, representative *Igl–Igh* translocation showing green and red signals juxtaposed on a dicentric chromosome (yellow arrow). Right, quantification of *Igh–Igl* translocations in day 4 control ($n = 2$) or *CD21-cre*, *Xrcc4*$^{c/-}$ ($n = 4$) B cells activated with anti-CD40 plus IL4, and analysed by metaphase FISH (see Supplementary Table 8). Data are presented as mean and s.d. **b**, PCR-isolated *Igh–Igl* translocation junctions from day-4-activated *CD21-cre*, *Xrcc4*$^{c/-}$ B cells ($n = 3$) (primers indicated by horizontal black arrows). Junctional sequences are shown in Supplementary Fig. 11. A vertical green arrow indicates breakpoints. For a given translocation, the same number is used to indicate the corresponding *Igh* and *Igl* breakpoints, with the *Igh* breakpoint denoted by a prime symbol.

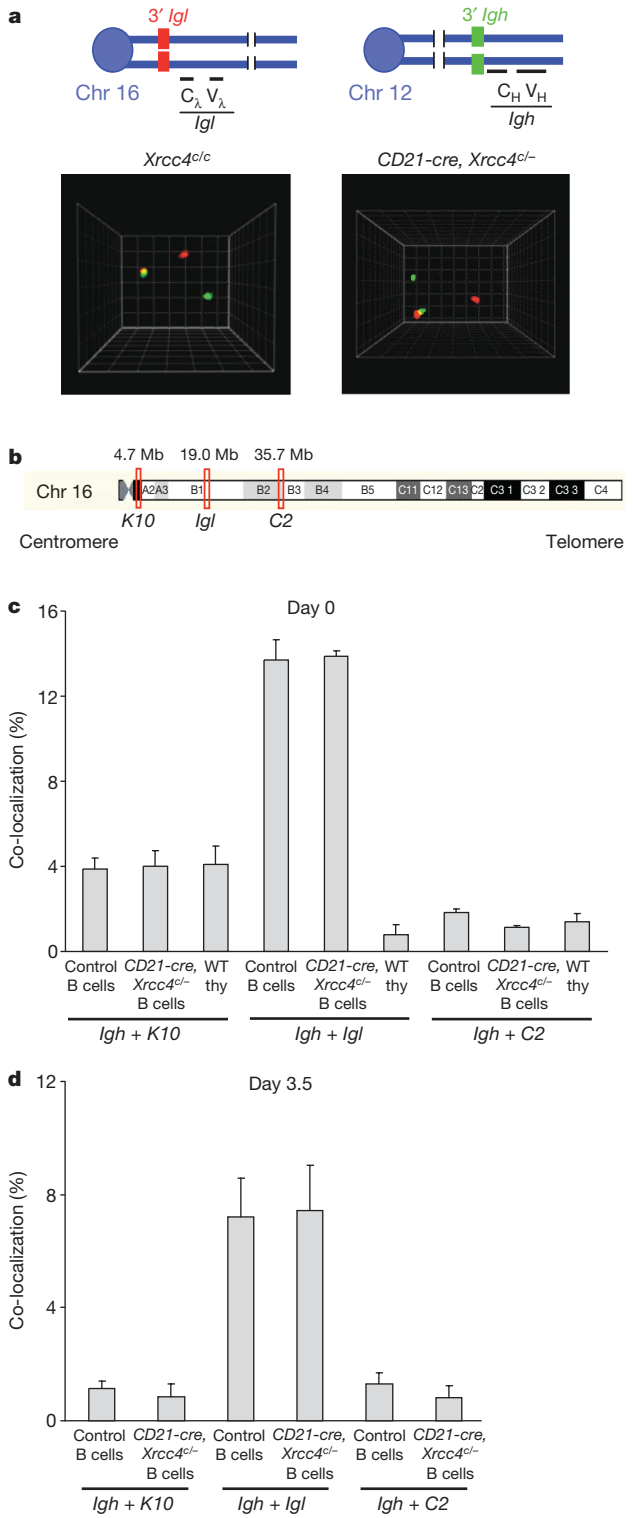


Figure 3 | Frequent cell-type and Igλ locus-specific *Igh* and *Igl* co-localization. **a**, Top, diagram showing 3′ *Igh* (green) and 3′ *Igl* (red) probes used for 3D interphase FISH. Bottom, representative co-localization of *Igh* and *Igl* in day 0 control and *CD21-cre*, *Xrcc4*$^{c/-}$ B cell interphase nuclei. **b**, Schematic map of *Igl*, *C2* and *K10* BAC probes on chromosome 16. **c**, Quantification of co-localization of *Igh* and *Igl*, *Igh* and *C2*, or *Igh* and *K10* loci in nuclei of day 0 control and *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells, and in nuclei of thymocytes (WT thy) (details in Supplementary Tables 9 and 11). **d**, Quantification of co-localization of *Igh* and *Igl*, *Igh* and *C2*, or *Igh* and *K10* loci in day-3.5-activated control or *CD21-cre*, *Xrcc4*$^{c/-}$ peripheral B cells (see Supplementary Table 10). At least three mice were analysed per data set; data are presented as mean and s.e.m.

conclude that *Igh* and *Igl* co-localize before and after DSB induction, and that XRCC4 deficiency does not alter this association. To assess cell-type specificity, we assayed wild-type thymocyte and ES cell interphase nuclei and found only low-level *Igh* and *Igl* co-localization (Fig. 3c and Supplementary Table 9). To examine specificity of the *Igh* and *Igl* association within chromosome 16, we tested co-localization of *Igh* with two control loci (C2 and K10), which map, respectively, about 15 Mb telomeric or centromeric to *Igl* (Fig. 3b). *Igh* and *C2* co-localization was at background levels in resting and activated B cells and thymocytes, whereas *Igh* and *K10* co-localization occurred at substantially lower levels than *Igh* and *Igl* co-localization (Fig. 3c and Supplementary Table 11). Therefore, *Igh* and *Igl* co-localization is cell-type specific and focal on chromosome 16 with respect to *Igl*. Notably, *Igh* and *Igk* also co-localize specifically and focally, at least with respect to *Igk*, in about 5% of splenic B cells (Supplementary Fig. 12 and Supplementary Table 12).

DSBs in c-myc are rate-limiting for translocations

Given that CXP tumours routinely have *Igh–c-myc* translocations[36], we tested for *Igh* and *c-myc* co-localization in B cell nuclei by 3D interphase FISH (Fig. 4b). Approximately, 4–6% of resting, 15-min activated and 3.5-day activated control or *CD21-cre*, *Xrcc4*$^{c/-}$ B cell nuclei had co-localized *Igh* and *c-myc* signals (Fig. 4b, c and Supplementary Table 13), which were specific as *Igh* and *c-myc* did not co-localize in ES cells (Fig. 4c and Supplementary Table 13). Although *c-myc* breaks and *Igh–c-myc* translocations were too infrequent to detect by FISH (Supplementary Table 14), PCR demonstrated an approximately fivefold increase in *Igh–c-myc* translocations in activated *CD21-cre*, *Xrcc4*$^{c/-}$ B cells compared to low ($<1\times10^{-6}$ per cell) control levels (Fig. 4a, Supplementary Fig. 13 and Supplementary Table 15). On the basis of frequent *Igh* breaks and *Igh–c-myc* co-localization, we hypothesized *c-myc* breaks to be rate-limiting for *Igh–c-myc* translocations. To test this, we introduced 25 tandemly arrayed ISceI endonuclease target sites[42] into the *c-myc* first intron to create the *c-myc*^{25IsceI} allele (Fig. 4d and Supplementary Fig. 14). Then, anti-CD40- and IL4-activated peripheral B cells heterozygous for the *c-myc*^{25IsceI} allele (*c-myc*$^{25ISceI/WT}$) or wild-type control B cells (*c-myc*$^{WT/WT}$) were infected with ISceI-expressing or control retrovirus[43] and assayed for *c-myc* breaks by metaphase FISH. Notably, *c-myc* chromosomal breaks occurred in approximately 10% of *c-myc*$^{25ISceI/WT}$ B cells infected with the ISceI virus, but were absent in the various control B cells (Fig. 4e and Supplementary Table 16). PCR quantification demonstrated that *Igh–c-myc* translocations in ISceI-virus-infected activated *c-myc*$^{25ISceI/WT}$ B cells were increased by at least 100-fold compared to control levels (Fig. 4a, Supplementary Fig. 15 and Supplementary Table 16).

Discussion

We show that some activated *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells contain characteristics of postulated 'editing and switching' CXP peripheral B cell lymphoma progenitors[36], including *Igk* deletions, aberrant *Igl* V(D)J recombination, *Igl* translocations, and aberrant *Igh* CSR associated with *Igh* translocations to *c-myc* or *Igl*. Moreover, our studies clearly show V(D)J recombination-related events in *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells; because they leave tell-tale RAG-dependent *Igl* breaks. Our cultured splenic B cells do not represent germinal centre B cells[44], and CXP tumour progenitors do not seem to be of germinal centre origin[36]. Therefore, we suggest that V(D)J recombination events in activated *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells and putative CXP lymphoma progenitors may represent peripheral editing mediated by low RAG expression, for example, as found in transitional B cells[8,45,46]. Although potential physiological roles for such a process are unknown, it may be relevant for peripheral B cells subjected to chronic activation, such as those in gut-associated lymphoid tissues where CXP tumours arise[36]. In this context, we find RAG-dependent *Igl* breaks in *CD21-cre*, *Xrcc4*$^{c/-}$ mesenteric lymph node B cells taken directly from mice (D.R.W. and F.W.A., unpublished data).

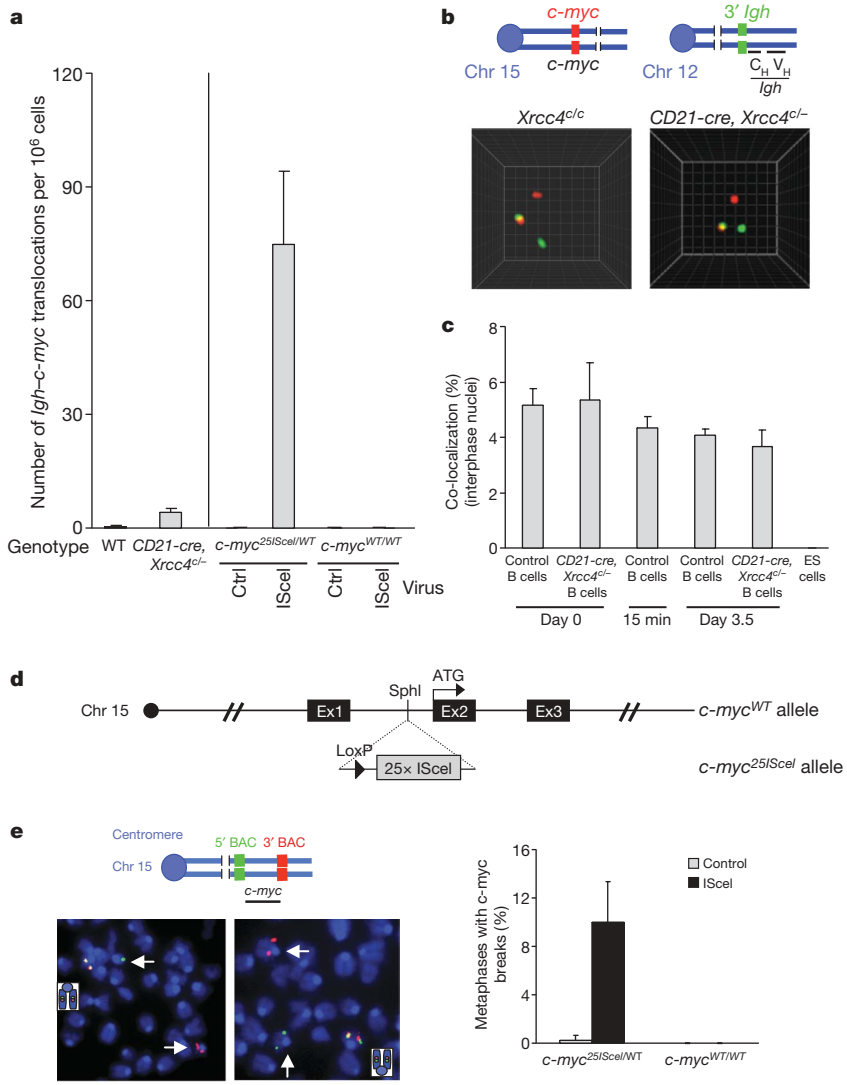


Figure 4 | DSBs in *c-myc* are rate-limiting for *Igh–c-myc* translocations in activated splenic B cells. a, Frequency of *Igh–c-myc* translocations from day 4 anti-CD40- and IL4-activated wild-type (WT) ($n = 4$) and *CD21-cre*, *Xrcc4*$^{c/-}$ ($n = 4$) splenic B cells, or B cells containing the wild-type *c-myc* allele (*c-myc*$^{WT/WT}$) ($n = 1$) or the modified *c-myc* allele containing 25 ISceI sites (*c-myc*$^{25ISceI/WT}$) ($n = 3$) with either control (Ctrl) or ISceI-expressing retrovirus (see Supplementary Figs 13 and 15). **b**, Top, schematic showing *c-myc* (red) probe on chromosome 15 and 3′ *Igh* (green) probe on chromosome 12. Bottom, representative images of *Igh* and *c-myc* co-localization in day 0 control and *CD21-cre*, *Xrcc4*$^{c/-}$ B cell interphase nuclei. **c**, Quantification of *Igh* and *c-myc* association by 3D interphase FISH in control and *CD21-cre*, *Xrcc4*$^{c/-}$ splenic B cells ($n = 3$), and ES cells ($n = 3$). Cells were analysed at the indicated time points before or after stimulation. **d**, Schematic showing the *c-myc*$^{WT/WT}$ allele and the modified *c-myc*^{25ISceI} allele. The ISceI array was used to increase cutting frequency. Ex, exon. **e**, Top left, diagram of *c-myc* FISH probes. Bottom left, representative *c-myc* abnormalities in anti-CD40- and IL-4-activated *c-myc*$^{25ISceI/WT}$ B cells infected with ISceI-expressing retrovirus, appearing as green and red signals on separate chromosome fragments (white arrows). Bottom right, quantification of *c-myc* breaks by metaphase FISH on day 4 anti-CD40- and IL4-activated B cells containing either *c-myc*$^{25ISceI/WT}$ ($n = 4$) or *c-myc*$^{WT/WT}$ ($n = 1$) alleles after infection with control or ISceI-expressing retrovirus. Data are presented as mean and s.d. (see Supplementary Table 16). High titre retrovirus infection seems to inhibit end-joining allowing break visualization (see Methods).

Our findings of RAG-initiated chromosomal breaks and translocations in XRCC4-deficient peripheral B cells raises the possibility that translocations in some human peripheral B cell lymphomas, such as follicular lymphomas, might be initiated by V(D)J recombination in the periphery[15,30]. They also demonstrate that AID and RAG can collaborate to generate frequent *Igh–Igl* translocations in peripheral *CD21-cre*, *Xrcc4*$^{c/-}$ B cells. Notably, these *Igh–Igl* translocations offer no obvious cellular selective advantage; therefore, their appearance as clonal translocations in CXP lymphomas may simply reflect the frequent occurrence of these translocations in tumour progenitors due to mechanistic factors that include the two loci being frequently broken and spatially proximal. In the latter context, we have shown that the co-localization of two loci on different chromosomes can be quite focal, implicating aspects of particular loci themselves, beyond broader chromosomal territories[47], as important factors in determining spatial proximity and translocation frequency. Furthermore, analyses of oncogenic translocations in NHEJ-deficient pro-B and B cell lymphomas[36,48] suggest alternative end-joining may be translocation-prone relative to NHEJ[49,50]. The high frequency of specific translocations catalysed by alternative end-joining in non-transformed *CD21-cre*, *Xrcc4*$^{c/-}$ B cells supports this notion.

METHODS SUMMARY

Generation of mouse strains. *CD21-cre*, *Xrcc4*$^{c/-}$ mice were generated as previously described[10] and crossed into AID-deficient mice[38] to generate *CD21-cre*, *Xrcc4*$^{c/-}$ *Aid*$^{-/-}$, or into mice carrying floxed *Rag2* alleles[40] to generate *CD21-cre*, *Xrcc4*$^{c/-}$ *Rag2*^{c} lines. We inserted a cassette containing 25 tandem ISceI target sites into the first intron of *c-myc* by gene targeting (see Methods). Mice were analysed as outlined in the text at 8–16 weeks of age. The Institutional Animal Care and Use Committee of Children's Hospital (Boston, Massachusetts) approved all animal work.

Splenic B cell purification, activation in culture, retroviral infection and CSR assays. CD43^{-} B cells were isolated from spleen, cultured, and assayed for CSR as previously described[10,22]. Cells were sampled on various days for DNA isolation, flow cytometry analysis and metaphase preparation. Retroviral infection was performed as previously described[43] (see Methods).

Two-colour FISH and telomere-FISH. Metaphase spreads from B cell cultures activated with anti-CD40 and IL4 were prepared and either two-colour FISH was used to detect *Igh*, *Igk*, *Igl* or *c-myc* chromosomal aberrations or telomere staining (T-FISH) was used to detect general aberrations as previously described[10]. FISH probes are detailed in Methods.

3D interphase FISH. 3D FISH was performed as described[32] (see Methods). Images of approximately 50 serial optical sections spaced by 0.2 μm were captured with Marianas spinning disk confocal microscope (×63) with a CCD detector (Intelligent Imaging Innovations) and analysed with Slidebook software (Intelligent Imaging Innovations).

PCR assay to detect *Igh–c-myc* or *Igh–Igl* translocations. *Igh–c-myc* translocation junctions were amplified by PCR from genomic DNA prepared from splenic B cells activated with anti-CD40 and IL4, using primers as described[26]. PCR products were run on agarose gels and hybridized with an internal *c-myc* oligonucleotide. *Igh–Igl* translocations were amplified using nested primers for *Sm* and *Jl*. PCR products were hybridized with *Jl* and *Igh* probes; the bands positive for both probes were cloned into the pGEM-T vector (Promega), sequenced and analysed using Lasergene software and the NCBI database. Primer sequences and PCR conditions are detailed in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 March; accepted 25 May 2009.

1. Jung, D. & Alt, F. W. Unraveling V(D)J recombination; insights into gene regulation. *Cell* **116,** 299–311 (2004).
2. Rooney, S., Chaudhuri, J. & Alt, F. W. The role of the non-homologous end-joining pathway in lymphocyte development. *Immunol. Rev.* **200,** 115–131 (2004).
3. Bassing, C. H., Swat, W. & Alt, F. W. The mechanism and regulation of chromosomal V(D)J recombination. *Cell* **109,** S45–S55 (2002).
4. Gorman, J. R. & Alt, F. W. Regulation of immunoglobulin light chain isotype expression. *Adv. Immunol.* **69,** 113–181 (1998).
5. Gay, D., Saunders, T., Camper, S. & Weigert, M. Receptor editing: an approach by autoreactive B cells to escape tolerance. *J. Exp. Med.* **177,** 999–1008 (1993).
6. Tiegs, S. L., Russell, D. M. & Nemazee, D. Receptor editing in self-reactive bone marrow B cells. *J. Exp. Med.* **177,** 1009–1020 (1993).
7. Nemazee, D. Receptor editing in lymphocyte development and central tolerance. *Nature Rev. Immunol.* **6,** 728–740 (2006).
8. Jankovic, M., Casellas, R., Yannoutsos, N., Wardemann, H. & Nussenzweig, M. C. RAGs and regulation of autoantibodies. *Annu. Rev. Immunol.* **22,** 485–501 (2004).
9. Chaudhuri, J. *et al.* Evolution of the immunoglobulin heavy chain class switch recombination mechanism. *Adv. Immunol.* **94,** 157–214 (2007).
10. Yan, C. T. *et al.* IgH class switching and translocations use a robust non-classical end-joining pathway. *Nature* **449,** 478–482 (2007).
11. Nemazee, D. & Weigert, M. Revising B cell receptors. *J. Exp. Med.* **191,** 1813–1817 (2000).
12. Seagal, J. & Melamed, D. Role of receptor revision in forming a B cell repertoire. *Clin. Immunol.* **105,** 1–8 (2002).
13. Wilson, P. C. *et al.* Receptor revision of immunoglobulin heavy chain variable region genes in normal human B lymphocytes. *J. Exp. Med.* **191,** 1881–1894 (2000).
14. Goossens, T., Brauninger, A., Klein, U., Kuppers, R. & Rajewsky, K. Receptor revision plays no major role in shaping the receptor repertoire of human memory B cells after the onset of somatic hypermutation. *Eur. J. Immunol.* **31,** 3638–3648 (2001).
15. Kuppers, R. & Dalla-Favera, R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene* **20,** 5580–5594 (2001).
16. Janz, S. Myc translocations in B cell and plasma cell neoplasms. *DNA Repair (Amst.)* **5,** 1213–1224 (2006).
17. Gostissa, M., Ranganath, S., Bianco, J. M. & Alt, F. W. Chromosomal location targets different MYC family gene members for oncogenic translocations. *Proc. Natl Acad. Sci. USA* **106,** 2265–2270 (2009).
18. Kozubek, S. *et al.* Distribution of *ABL* and *BCR* genes in cell nuclei of normal and irradiated lymphocytes. *Blood* **89,** 4537–4545 (1997).
19. Neves, H., Ramos, C., da Silva, M. G., Parreira, A. & Parreira, L. The nuclear topography of ABL, BCR, PML, and RARα genes: evidence for gene proximity in specific phases of the cell cycle and stages of hematopoietic differentiation. *Blood* **93,** 1197–1207 (1999).
20. Nikiforova, M. N. *et al.* Proximity of chromosomal loci that participate in radiation-induced rearrangements in human cells. *Science* **290,** 138–141 (2000).
21. Roix, J. J., McQueen, P. G., Munson, P. J., Parada, L. A. & Misteli, T. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nature Genet.* **34,** 287–291 (2003).
22. Osborne, C. S. *et al.* Myc dynamically and preferentially relocates to a transcription factory occupied by IgH. *PLoS Biol.* **5,** e192 (2007).
23. Meaburn, K. J., Misteli, T. & Soutoglou, E. Spatial genome organization in the formation of chromosomal translocations. *Semin. Cancer Biol.* **17,** 80–90 (2007).
24. Tsai, A. G. *et al.* Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. *Cell* **135,** 1130–1142 (2008).
25. Mahowald, G. K., Baron, J. M. & Sleckman, B. P. Collateral damage from antigen receptor gene diversification. *Cell* **135,** 1009–1012 (2008).
26. Ramiro, A. R. *et al.* AID is required for c-myc/IgH chromosome translocations *in vivo*. *Cell* **118,** 431–438 (2004).
27. Ramiro, A. *et al.* The role of activation-induced deaminase in antibody diversification and chromosome translocations. *Adv. Immunol.* **94,** 75–107 (2007).
28. Kovalchuk, A. L. *et al.* AID-deficient Bcl-xL transgenic mice develop delayed atypical plasma cell tumors with unusual Ig/Myc chromosomal rearrangements. *J. Exp. Med.* **204,** 2989–3001 (2007).
29. Robbiani, D. F. *et al.* AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. *Cell* **135,** 1028–1038 (2008).
30. Jager, U. *et al.* Follicular lymphomas' BCL-2/IgH junctions contain templated nucleotide insertions: novel insights into the mechanism of t(14;18) translocation. *Blood* **95,** 3520–3529 (2000).
31. Lieber, M. R., Yu, K. & Raghavan, S. C. Roles of nonhomologous DNA end joining, V(D)J recombination, and class switch recombination in chromosomal translocations. *DNA Repair (Amst.)* **5,** 1234–1245 (2006).
32. Callen, E. *et al.* ATM prevents the persistence and propagation of chromosome breaks in lymphocytes. *Cell* **130,** 63–75 (2007).
33. Li, Z. *et al.* The *XRCC4* gene encodes a novel protein involved in DNA double-strand break repair and V(D)J recombination. *Cell* **83,** 1079–1089 (1995).
34. Gao, Y. *et al.* A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis. *Cell* **95,** 891–902 (1998).
35. Soulas-Sprauel, P. *et al.* Role for DNA repair factor XRCC4 in immunoglobulin class switch recombination. *J. Exp. Med.* **204,** 1717–1727 (2007).
36. Wang, J. H. *et al.* Oncogenic transformation in the absence of Xrcc4 targets peripheral B cells that have undergone editing and switching. *J. Exp. Med.* **205,** 3079–3090 (2008).
37. Kraus, M., Alimzhanov, M. B., Rajewsky, N. & Rajewsky, K. Survival of resting mature B lymphocytes depends on BCR signaling via the Igα/β heterodimer. *Cell* **117,** 787–800 (2004).
38. Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102,** 553–563 (2000).
39. Gartner, F., Alt, F. W., Monroe, R. J. & Seidl, K. J. Antigen-independent appearance of recombination activating gene (RAG)-positive bone marrow B cells in the spleens of immunized mice. *J. Exp. Med.* **192,** 1745–1754 (2000).
40. Hao, Z. & Rajewsky, K. Homeostasis of peripheral B cells in the absence of B cell influx from the bone marrow. *J. Exp. Med.* **194,** 1151–1164 (2001).
41. Schrader, C. E., Linehan, E. K., Mochegova, S. N., Woodland, R. T. & Stavnezer, J. Inducible DNA breaks in Ig S regions are dependent on AID and UNG. *J. Exp. Med.* **202,** 561–568 (2005).
42. Plessis, A., Perrin, A., Haber, J. E. & Dujon, B. Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. *Genetics* **130,** 451–460 (1992).
43. Zarrin, A. A. *et al.* Antibody class switching mediated by yeast endonuclease-generated DNA breaks. *Science* **315,** 377–381 (2007).
44. Lahvis, G. P. & Cerny, J. Induction of germinal center B cell markers *in vitro* by activated $CD4^+$ T lymphocytes: the role of CD40 ligand, soluble factors, and B cell antigen receptor cross-linking. *J. Immunol.* **159,** 1783–1793 (1997).
45. Monroe, R. J. *et al.* RAG2:GFP knockin mice reveal novel aspects of RAG2 expression in primary and peripheral lymphoid tissues. *Immunity* **11,** 201–212 (1999).
46. Yu, W. *et al.* Continued RAG expression in late stages of B cell development and no apparent re-induction after immunization. *Nature* **400,** 682–687 (1999).
47. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Rev. Genet.* **2,** 292–301 (2001).
48. Zhu, C. *et al.* Unrepaired DNA breaks in p53-deficient cells lead to oncogenic gene amplification subsequent to translocations. *Cell* **109,** 811–821 (2002).
49. Roth, D. B. Amplifying mechanisms of lymphomagenesis. *Mol. Cell* **10,** 1–2 (2002).
50. McVey, M. & Lee, S. E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet.* **24,** 529–538 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Alt laboratory members for discussions, and Y. L. Chen, J. M. Bianco and M. Moghimi for technical assistance. This work was supported by the National Insitutes of Health (NIH) grant 5P01CA92625 and a Leukemia and Lymphoma Society of America (LLS) SCORE grant to F.W.A. and K.R. M.G. is and J.H.W. was a Special Fellow of the LLS. J.H.W. and D.R.W. are supported by an NIH training grant and C.T.Y. was supported by an NCI training grant. A.N. is supported by the Intramural Research program of the NIH, NCI, Center for Cancer Research. F.W.A. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions F.W.A., J.H.W., M.G. and C.T.Y. planned studies and interpreted data. J.H.W. performed most experiments, including mouse breeding, B cell studies, FISH, and *Igh* and *Igl* PCR studies. C.T.Y. bred mice and performed B cell analyses. M.G. generated and analysed *c-myc*$^{25IsceI/WT}$ mice and performed FISH and *Igh–c-myc* translocation studies. P.G., T.H., and E.H. provided technical assistance. S.D. and A.N. provided expertise in 3D interphase FISH. A.A.Z. generated the 25 IsceI array. D.R.W. performed RAG expression studies and mesenteric lymph node B cell analyses. K.R. provided RAG conditional knockout mice and helped interpret data. F.W.A., J.H.W. and M.G. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.W.A. (alt@enders.tch.harvard.edu).

METHODS

Gene targeting of *c-myc*^{25ISceI} allele. To generate the targeting construct, the ISceI site cassette was inserted into the pLNTK targeting vector and flanked with a 4.6-kb SphI–SphI 5′ homology arm encompassing *c-myc* exon 1, and a 2.6-kb SphI–SphI 3′ homology arm containing *c-myc* exons 2 and 3. We electroporated the targeting construct into TC1 (129/Sv) ES cells, and screened EcoRI-digested DNA to identify potentially targeted ES cell clones by Southern blotting with a 5′ probe consisting of the 1.5-kb XbaI–XbaI fragment upstream of *c-myc* exon 1, and a 3′ probe consisting of the 1.5-kb XhoI–KpnI fragment downstream of *c-myc* exon 3. We deleted the neomycin cassette from targeted ES clones by infection with adenovirus expressing Cre-recombinase and used those to inject for germline transmission to generate *c-myc*$^{25IsceI/WT}$ mice.

Retroviral infection and CSR assays. Empty pMX or pMX-IsceI vectors were transfected into Phoenix packaging cells. Viral supernatants were collected at 48 and 72 h after transfection, and used for infecting B cells stimulated with anti-CD40 and IL4 at days 1 and 2 of stimulation. Cells were processed at day 4 of stimulation. Infection efficiency ranged from 70% to 90% as measured by IRES–GFP expression from the retroviral vector by flow cytometry. High titre retroviral infection of activated wild-type B cells leads to inhibition of end joining as evidenced by increased accumulation of AID-initiated *Igh* chromosome breaks (M.G. and F.W.A., unpublished data). This effect facilitates visualization of ISceI-induced *c-myc* breaks in *c-myc*$^{25ISceI/WT}$ wild-type B cells.

Cytogenetic analysis: two colour FISH and FISH probes. The FISH probes were as follows: a BAC that covered the 3′ region of the *Igh* locus, encompassing the 3′ *Igh* enhancer and 100 kb of downstream sequences (3′ *Igh* BAC) and a BAC just upstream of the *Igh* V_H region (5′ *Igh* BAC) as described previously[10]. All BACs outlined below were obtained from the BACPAC CHORI database. BACs for *Igl* regions are RP23-382P9 (5′ *Igl*) and RP23-374P12 (3′ *Igl*), and BACs for *Igk* regions are RP23-84F6 (5′ *Igk*), RP24-279F16 (*Igk*) and RP23-64I9 (3′ *Igk*). BACs for *c-myc* regions are RP23-307D14 that contains the *c-myc* gene for interphase FISH, RP24-434C10 (5′ *c-myc*) and RP23-113O21 (3′ *c-myc*). In all FISH experiments, intact loci show co-localization of the red and green probes, whereas split red and green signals are scored as broken loci. Broken loci can be free, with the centromeric and telomeric portion of the locus either present in the metaphases or lost. Alternatively, broken loci (both the centromeric and telomeric parts or only one of them) can be involved in translocations with other chromosomes. For *Igh–Igl* translocation FISH, representative *Igh–Igl* translocation shows green and red signals juxtaposed on a dicentric chromosome. Also visible are a normal chromosome 12 with a red signal at the telomeric end and a normal chromosome 16 with a green signal close to the centromere. Whole chromosome paint specific for mouse chromosomes 6, 12 and 16 were used according to the manufacturer's instructions (Applied Spectral Imaging). Images of metaphases were captured with BX61 Microscope (Olympus) equipped with a motorized automatic stage, a cooled-CCD camera and an interferometer (Applied Spectral Imaging). A ×60 objective lens was used. Analysis was performed with the ScanView software (Applied Spectral Imaging).

3D interphase FISH. B cells, thymocytes or ES cells were transferred to a glass slide coated with Cell Tak (Becton Dickinson), fixed for 10 min with 4% paraformaldehyde, permeabilized in 0.1% saponin, 0.1% Triton X-100 in PBS for 10 min, and immersed in liquid nitrogen for three freeze/thaw cycles. Slides were incubated in 0.1 M HCl, blocked with 3% BSA and 100 mg ml^{-1} RNase A, and re-permeabilized in 0.5% saponin, 0.1% Triton X-100 in PBS. Genomic DNA was denatured at 73 °C and hybridized with FISH probes labelled with either biotin or digoxigenin (Roche). Detecting antibodies were streptavidin-Alexa 488 (Molecular Probes) and mouse anti-Dig-cy3 (Jackson Immuno Research laboratory) used at 1:200 dilution. We could not perform *Igk* locus assays because the large size of *Igk* (greater than 3 Mb) obviates chromosomal break measurements by interphase FISH.

Primers and PCR conditions for *Igh–Igl* translocations. Primers for *Jl1* regions: first round primer, 5′-ATTGAATTTCTGTGGCCATAACTCC-3′, second round primer, 5′-ACTCCCCTCTGTTTTTACTCCTTATTA-3′. Primers for *Jl2* regions: first round primer, 5′- AGTTATTTCCTCTTTGCTCCTGTGGT GAT-3′, second round primer, 5′-CTACTCTGGGGTTGGTGCCTACTG-3′. Primers for *Igh Sm* regions: first round primer, 5′-GATCCCTCTACACCC TGCTATTTC-3′, second round primer, 5′-CACCCTGCTATTTCCTTGTT GCTACTCTTA-3′. Probes: for *Jl1* 5′-GCCTCTGTGCTGAATGTTCTGTGC TC-3′, for *Jl2* 5′-TCAGGAGGTAGAGCACAGCAGCCACTT-3′, for *Igh* 5′-CC TGGTATACAGGACGAAACTGCAGCAG-3′. The conditions for the first round PCR were 94 °C 3 min, 94 °C 15 s, 58.2 °C 15 s, 68 °C 7 min plus 20 s per cycle, 25 cycles, and 68 °C 5 min for final extension. Conditions for the second round PCR were 94 °C 3 min, 94 °C 15 s, 58.2 °C 15 s for *Igh* and *Jl1*, or 60.3 °C, 15 s for *Igh* and *Jl2*, 68 °C 7 min, 25 cycles, 68 °C 5 min for final extension.

Primers and PCR conditions for *Igh–c-myc* translocations. Primers for *c-myc*^{WT}: first round primer, 5′-GGGGAGGGGGTGTCAAATAATAAGA-3′, second round primer, 5′-GACACCTCCCTTCTACACTCTAAACCG-3′. Primers for *c-myc*^{25ISceI}: first round primer, 5′-GCTCTGCTGTTGCTGGTGAT AGAA-3′, second round primer, 5′-ATAGGGCTGTACGGAGTCGTAGTC-3′. Primers for *Igh Sm* region: first round primer, 5′-TGAGGACCAGAGAGGG ATAAAAGAGAA-3′, second round primer, 5′-CACCCTGCTATTTCCTTGT TGCTAC-3′. The conditions for first round PCR were 94 °C 3 min, 94 °C 15 s, 62 °C 15 s, 68 °C 7 min plus 20 s per cycle, 25 cycles, 68 °C 5 min for final extension. Conditions for the second round PCR were 94 °C 3 min, 94 °C 15 s, 62 °C 15 s, 68 °C 7 min, 25 cycles, 68 °C 5 min for final extension.

LETTERS

Type IIn supernovae at redshift $z \approx 2$ from archival data

Jeff Cooke[1], Mark Sullivan[2], Elizabeth J. Barton[1], James S. Bullock[1], Ray G. Carlberg[3], Avishay Gal-Yam[4] & Erik Tollerud[1]

Supernovae have been confirmed to redshift $z \approx 1.7$ (refs 1, 2) for type Ia (thermonuclear detonation of a white dwarf) and to $z \approx 0.7$ (refs 1, 3–5) for type II (collapse of the core of the star). The subclass type IIn (ref. 6) supernovae are luminous[7–9] core-collapse explosions of massive stars[8–11] and, unlike other types, are very bright in the ultraviolet[12–15], which should enable them to be found optically at redshifts $z \approx 2$ and higher[14,16]. In addition, the interaction of the ejecta with circumstellar material creates strong, long-lived emission lines that allow spectroscopic confirmation of many events of this type at $z \approx 2$ for 3–5 years after explosion (ref. 14). Here we report three spectroscopically confirmed type IIn supernovae, at redshifts $z = 0.808$, 2.013 and 2.357, detected in archival data using a method[14] designed to exploit these properties at $z \approx 2$. Type IIn supernovae directly probe the formation of massive stars at high redshift. The number found to date is consistent with the expectations of a locally measured[17] stellar initial mass function, but not with an evolving initial mass function proposed[18–20] to explain independent observations at low and high redshift.

The three $z \approx 2$ type IIn supernovae are detected in the Deep component of the Canada–France–Hawaii Telescope Legacy Survey (CFHT-LS), which consists of four fields, each of area one degree-squared, imaged over five years in five (u*, g′, r′, i′, z′) filters. Our approach is to monitor $z \approx 2$ galaxies over multiple years and search for flux variations that meet criteria for high-redshift type IIn supernovae. Galaxies are identified using efficient colour-selection techniques[21–23] tailored to the CFHT-LS and spectroscopically tested using the Vi(r)mos-VLT Deep Survey[24]. The nightly images from a given year (a 5–6 month 'season' of observations) are combined into 'seasonal-stacked' images for each filter. The stacked images are sensitive to the slow (~3–5 months at $z \approx 2$) photometric evolution of high-redshift type IIn supernovae as a result of cosmological time dilation, and to events ~6 times fainter than those detected by the Supernova Legacy Survey that searches the same high-quality data using conventional techniques[25]. Finally, we use the 'seasonal-stacked' images for candidate inspection and the nightly exposures to construct high-resolution light curves (see Supplementary Information for more details).

The confirmed $z \approx 2$ type IIn supernovae are the first three supernova candidates detected in the two fields over four seasons (2003–06) analysed to date. Candidates must meet conservative criteria that include (1) detection of $\geq 3\sigma$ in the g′-, r′- and i′-band 'seasonal-stacked' images, (2) flux variation in only one season (the first and last seasons are disregarded for this reason), (3) clean point-source detections in the subtracted images, (4) a priority for events with host galaxy centroid offsets, and (5) g′r′i′ light curves that exhibit flux rise times and decay rates consistent with supernova profiles. The criteria are designed to prevent misidentification of active galactic nuclei (AGN; accreting supermassive black holes at the centres of galaxies) and other contaminants that can mimic supernova events to a certain extent.

The g′r′i′ light curves for the three $z \approx 2$ supernovae are presented in Fig. 1 and are photometrically consistent with type IIn supernova behaviour. We confirmed the supernova redshifts (Table 1) from deep spectroscopy obtained using the 10-m telescopes at the W. M. Keck Observatory equipped with the Deep Imaging Multi-Object Spectrograph[26] on 30 September 2008 and 1 October 2008 and with the Low Resolution Imaging Spectrometer[27,28] on 25 January 2009. The ability to spectroscopically confirm the events as type IIn supernovae via the detection of strong, long-lived emission lines differs dramatically from other supernova types that require rapid follow-up to obtain spectral classifications of quickly fading continua.

We detected late-time supernova emission to $>3\sigma$ significance in the combined 6,000-s exposure of SN 234161 (365 days old, rest-frame; Fig. 2). The features are very similar to those seen at low redshift, and include ultraviolet shock ionization emission lines (such as semi-forbidden transitions N IV] and N III] that are extremely rare in AGN), and emission-line strength ratio values (for example, C IV/C III]) that are not theoretically predicted or observed in AGN. In addition, the combined 5,400-s exposure of SN 19941 (345 days old,

Table 1 | High-redshift supernovae and host galaxy properties

SN	RA* (J2000)	Dec.† (J2000)	Host m_r (mag)	Date of outburst	SN m_r (mag)	SN M_{UV} (mag)	Separation (kpc)	Redshift z
19941	02 24 11.147	−04 57 58.41	25.16 ± 0.07	24 Nov. 2005	$25.6^{+0.2}_{-0.3}$	$-19.5^{+0.2}_{-0.3}$	1.1 ± 0.8	2.357 ± 0.002
234161	02 24 33.271	−04 26 31.04	24.93 ± 0.07	27 Sep. 2005	$25.2^{+0.2}_{-0.2}$	$-19.6^{+0.2}_{-0.2}$	2.8 ± 0.8	2.013 ± 0.002
219241	22 14 47.351	−17 46 11.62	24.12 ± 0.04	8 Jul. 2004	$24.7^{+0.1}_{-0.1}$	$-18.2^{+0.1}_{-0.1}$	2.0 ± 0.7	0.808 ± 0.001

Both the dates of outburst and the supernova (SN) magnitudes are determined from the light curves. Supernova apparent r′-band magnitudes (m_r) are at peak brightness and corresponding absolute ultraviolet magnitudes (M_{UV}) are estimated at the r′-band effective rest-frame wavelengths. Separations are measured between the centroids of the supernovae and their host galaxies and are accurate to ± 0.52 pixels, which corresponds to a physical separation of 0.7–0.8 kpc (see Supplementary Information).
* RA, right ascension (in h min s).
† Dec., declination (in ° ′ ″).

[1]Center for Cosmology, Department of Physics and Astronomy, University of California, Irvine, California 92697-4574, USA. [2]Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK. [3]Department of Astronomy and Astrophysics, University of Toronto, Toronto, Ontario M5S 3H4, Canada. [4]Benoziyo Center for Astrophysics, Weizmann Institute of Science, 76100 Rehovot, Israel.

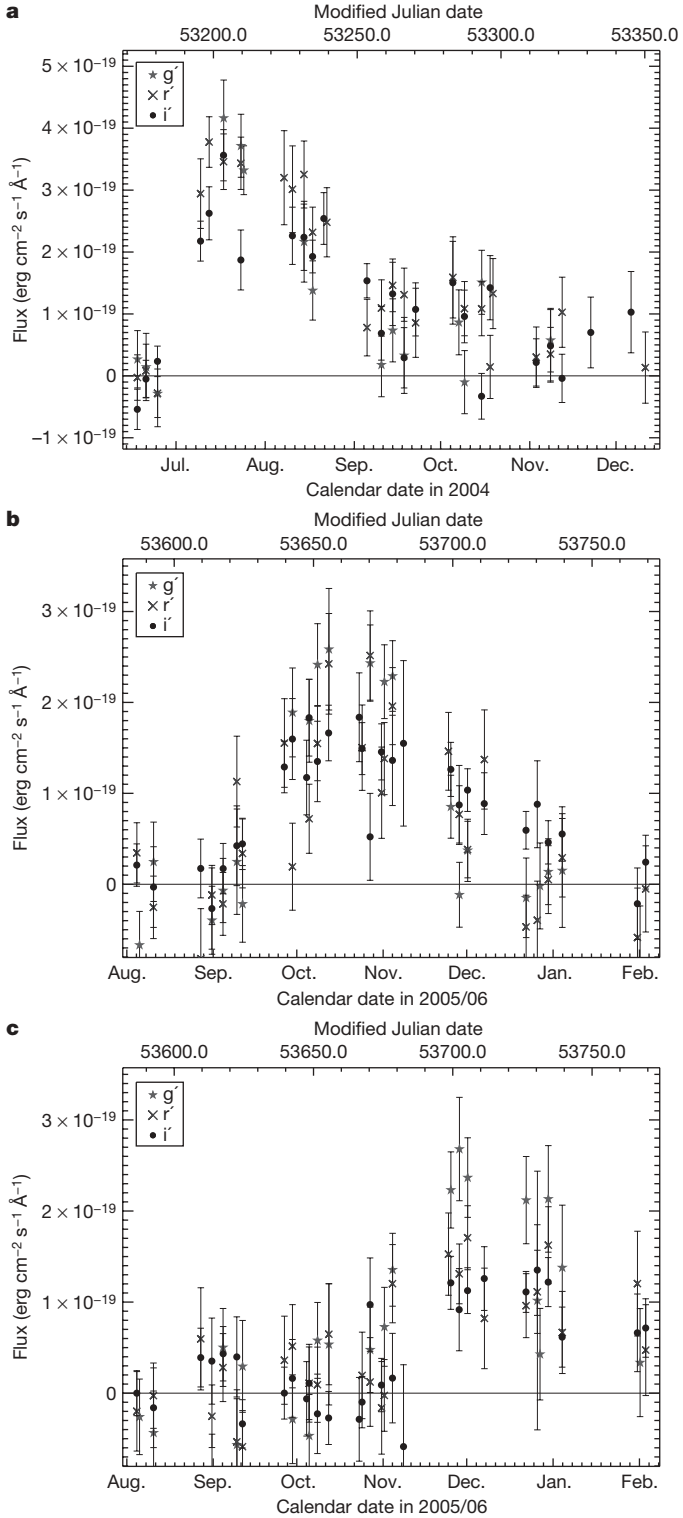


Figure 1 | Multi-colour light curves of the three high-redshift supernovae. **a**, SN 219241 ($z = 0.808$); **b**, SN 234161 ($z = 2.013$); and **c**, SN 19941 ($z = 2.357$). The flux and 1σ uncertainties for the g′r′i′ optical filters shown here probe the rest-frame ultraviolet flux for these events. The seasonal integrated flux from each supernova is detected at 3–8.4σ over the host galaxy flux in each filter. Because type Ia supernovae exhibit very little flux shortward of ~300 nm as a result of efficient Fe II scattering of ultraviolet photons[29], this classification is ruled out. We find that the ultraviolet luminosity and evolution of the supernovae in each filter are most consistent with type IIn events (see Supplementary Information).

rest-frame) exhibits strong blueshifted Si IV and weaker C IV emission attributed to the supernova. We did not detect significant ($>3\sigma$) supernova emission in the spectrum of SN 219241. This is probably a consequence of the supernova age (855 days old, rest-frame), placing it at an epoch when the emission lines are expected to fade to near or below the spectroscopic threshold of the shorter combined 3,600-s exposure[14]. The detection and eventual decay of type IIn supernova emission lines verify photometric classification, and confirm the ability to study supernova energies and chemistry on an individual basis to high redshift. (See Supplementary Information for spectroscopic and line-emission details.)

Progenitors of type IIn supernovae are believed to be massive stars[8–11] that sample the high-mass end of the stellar initial mass function (IMF) of galaxies. Because our method colour-selects a well-controlled population of galaxies over a well-defined volume, a small number of type IIn supernovae can not only give the high-redshift rate of these supernovae, but also provide the first direct probe of the high-redshift IMF. For example, an evolving IMF model invoked recently to reconcile indirect high- and low-redshift observations[18–20] predicts a greater number (by factor of ~3) of $z \approx 2$ type IIn supernovae than a locally measured IMF[17]. Computing the expectations for the volume

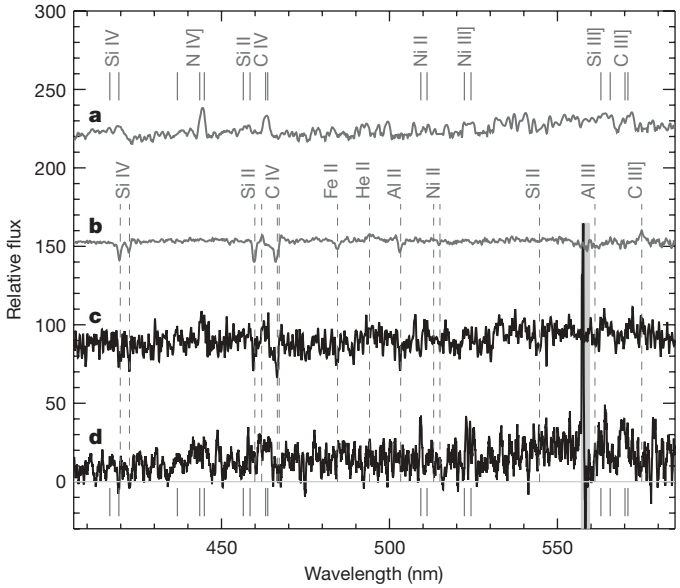


Figure 2 | Four spectra with vertical offsets to help illustrate the detection of line-emission from SN 234161. **a**, The ultraviolet spectrum of the low-redshift SN 1998S[13] (day 485) shifted and flux corrected to $z = 2.013 - 2{,}500\ \mathrm{km\ s^{-1}}$ ($2{,}500\ \mathrm{km\ s^{-1}}$ corresponds to $\Delta z = -0.026$; best fit match to SN 234161 emission). The short solid vertical lines indicate the positions of expected type IIn supernova emission-line features. **b**, A composite $z \approx 3$ galaxy spectrum[30] representative of typical $z \approx 2$–3 galaxies using our colour selection. Dashed vertical lines indicate the expected positions of interstellar ultraviolet features. **c**, Convolution of the composite galaxy and the SN 1998S spectra, with Gaussian noise added consistent with that of the data. **d**, The data for SN 234161 (day 365). The thick grey vertical line marks the position of a bright night sky emission line that is difficult to subtract cleanly from the faint spectrum. The effect of the supernova emission on the convolved spectrum includes the Si IV 139.4, 140.3 nm profile, excess flux from N IV] 146.1, 148.3, 148.8 and N III] 174.1, 174.7, 175.3 nm near 440 and 520 nm, respectively, excess flux and profile of C IV 154.8, 155.0 nm near 460 nm, and the rise in the continuum beyond ~530 nm. Comparison of the data to the convolved spectrum shows that SN 234161 is consistent with an average $z \approx 2$–3 galaxy experiencing a SN 1998S-like event with emission-line peaks blue-shifted $2{,}500\ \mathrm{km\ s^{-1}}$ (low-redshift type IIn supernovae also exhibit blue-shifted emission peaks). In addition, SN 234161 exhibits strong C III] 190.7, 190.9 nm supernova emission near 570 nm. We remark that low-redshift supernovae SN 1998S and type IIn SN 1995N[12] exhibited C III] emission of similar relative strength to SN 234161 at earlier (day 238) and later (day 943) times, respectively.

analysed to date, the confirmed type IIn supernovae presented here are consistent with the relative number predicted using a static local IMF. A relaxation of our conservative supernova criteria and a search for supernovae at greater radii from their host centroids in the complete CFHT-LS data set will test the validity the of this result. Finally, the method presented here provides a means to identify ~40,000 type IIn supernovae at $z \approx 2$ and to detect events to $z \approx 6$ over the next ten years with 8-m-class deep synoptic optical campaigns, some of which are currently under way and some of which are soon to begin. This cannot be done with any other supernova type or by conventional search practices. As a result, the exceptional properties of type IIn supernovae should enable a seamless study of stellar and galactic processes, ranging from the local Universe to a time shortly after the formation of the first stars.

Received 3 February; accepted 21 April 2009.

1. Poznanski, D. *et al.* Supernovae in the Subaru deep field: an initial sample and type Ia rate out to redshift 1.6. *Mon. Not. R. Astron. Soc.* **382,** 1169–1186 (2007).
2. Riess, A. G. *et al.* The farthest known supernova: support for an accelerating Universe and a glimpse of the epoch of deceleration. *Astrophys. J.* **560,** 49–71 (2001).
3. Botticella, M. T. *et al.* Supernova rates from the Southern inTermediate Redshift ESO Supernova Search (STRESS). *Astron. Astrophys.* **479,** 49–68 (2008).
4. Della Valle, M. *et al.* Hypernova signatures in the late rebrightening of GRB 050525A. *Astrophys. J.* **642,** 103–106 (2006).
5. Soderberg, A. M. *et al.* An HST study of the supernovae accompanying GRB 040924 and GRB 041006. *Astrophys. J.* **636,** 391–399 (2006).
6. Schlegel, E. M. A new subclass of type II supernovae? *Mon. Not. R. Astron. Soc.* **244,** 269–271 (1990).
7. Richardson, D. *et al.* A comparative study of the absolute magnitude distributions of supernovae. *Astron. J.* **123,** 745–752 (2002).
8. Smith, N. *et al.* SN 2006gy: discovery of the most luminous supernova ever recorded, powered by the death of an extremely massive star like η Carinae. *Astrophys. J.* **666,** 1116–1128 (2007).
9. Smith, N. *et al.* SN 2006tf: precursor eruptions and the optically thick regime of extremely luminous type IIn supernovae. *Astrophys. J.* **686,** 467–484 (2008).
10. Kotak, R. & Vink, J. S. Luminous blue variables as the progenitors of supernovae with quasiperiodic radio modulations. *Astron. Astrophys.* **460,** 5–8 (2006).
11. Gal-Yam, A. *et al.* On the progenitor of SN 2005gl and the nature of type IIn supernovae. *Astrophys. J.* **656,** 372–381 (2007).
12. Fransson, C. *et al.* Optical and ultraviolet spectroscopy of SN 1995N: evidence for strong circumstellar interaction. *Astrophys. J.* **572,** 350–370 (2002).
13. Fransson, C. *et al.* Hubble Space Telescope and ground-based observations of SN 1993J and SN 1998S: CNO processing in the progenitors. *Astrophys. J.* **622,** 991–1007 (2005).
14. Cooke, J. Detecting $z > 2$ type IIn supernovae. *Astrophys. J.* **677,** 137–145 (2008).
15. Brown, P. *et al.* Ultraviolet light curves of supernovae with Swift UVOT. Preprint at ⟨http://arXiv.org/abs/0803.1265⟩ (2008).
16. Dahlén, T. & Fransson, C. Rates and redshift distributions of high-z supernovae. *Astron. Astrophys.* **350,** 345–367 (1999).
17. Salpeter, E. E. The luminosity function and stellar evolution. *Astrophys. J.* **121,** 161–167 (1955).
18. van Dokkum, P. G. Evidence of cosmic evolution of the stellar initial mass function. *Astrophys. J.* **674,** 29–50 (2008).
19. Chary, R.-R. The stellar initial mass function at the epoch of reionization. *Astrophys. J.* **680,** 32–40 (2008).
20. Davé, R. The galaxy stellar mass-star formation rate relation: evidence for an evolving initial mass function. *Mon. Not. R. Astron. Soc.* **385,** 147–160 (2008).
21. Steidel, C. C. *et al.* Lyman break galaxies at redshift $z \sim 3$: survey description and full data set. *Astrophys. J.* **592,** 728–754 (2003).
22. Steidel, C. C. *et al.* A survey of star-forming galaxies in the $1.4 < z$ 2.5 redshift desert: overview. *Astrophys. J.* **604,** 534–550 (2004).
23. Cooke, J. *et al.* Survey for galaxies associated with $z \sim 3$ damped Ly-α systems. I. Spectroscopic calibration of *u*′BVRI photometric selection. *Astrophys. J.* **621,** 596–614 (2005).
24. Le Fèvre, O. *et al.* Virmos-VLT deep survey (VVDS). *Proc. SPIE* **4834,** 173–182 (2003).
25. Neill, J. D. *et al.* The type Ia supernova rate at $z \approx 0.5$ from the supernova legacy survey. *Astron. J.* **132,** 1126–1145 (2006).
26. Faber, S. M. *et al.* The DEIMOS spectrograph for the Keck II telescope: integration and testing. *Proc. SPIE* **4841,** 1657–1669 (2003).
27. Oke, J. B. *et al.* The Keck low-resolution imaging spectrometer. *Publ. Astron. Soc. Pacif.* **107,** 375–385 (1995).
28. McCarthy, J. K. *et al.* Blue channel of the Keck low-resolution imaging spectrometer. *Proc. SPIE* **3355,** 81–91 (1998).
29. Riess, A. G. *et al.* Identification of type Ia supernovae at redshift 1.3 and beyond with the Advanced Camera for Surveys on the Hubble Space Telescope. *Astrophys. J.* **600,** L163–L166 (2004).
30. Shapley, A. E. *et al.* Rest-frame ultraviolet spectra of $z \sim 3$ Lyman break galaxies. *Astrophys. J.* **558,** 65–89 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements J.C. thanks D. Leonard for discussions. This work was made possible by support from the Gary McCue postdoctoral fellowship and the Center for Cosmology at the University of California, Irvine. We acknowledge support from NSERC and the Royal Society. The analysis pipeline used to reduce the DEIMOS data was developed at UC Berkeley with support from the NSF (grant AST-0071048). The CFHT-LS relies on observations with MegaCam, a joint project of CFHT and CEA/DAPNIA, at the CFHT. We used data products from the Canadian Astronomy Data Centre as part of the CFHT-LS. Both CFHT and the W. M. Keck Observatory are located near the summit of Mauna Kea, Hawai'i. We acknowledge the cultural role that the summit of Mauna Kea has within the indigenous Hawaiian community.

Author Contributions J.C. led the survey design, and performed observations, data reduction and analysis, candidate selection, and data interpretation. M.S. was responsible for the data reduction and analysis of the CFHT-LS yearly stacked images, the supernova light curves, interpretation and key discussions. E.J.B. performed observations and provided observing time, discussions, and input regarding survey design, data analysis and interpretation. J.S.B. provided observing time, performed observations, contributed discussions, and helped in manuscript preparation. R.G.C. was one of the proponents of the CFHT-LS and became the Canadian PI, assembling the team responsible for much of the real-time operation and analysis of the Supernova Legacy Survey. A.G.-Y. provided advice about the analysis and helped prepare the manuscript. E.T. performed spectroscopic observations for two of the $z \approx 2$ supernovae.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.C. (cooke@uci.edu).

Vol 460|9 July 2009|doi:10.1038/nature08121

LETTERS

Demonstration of two-qubit algorithms with a superconducting quantum processor

L. DiCarlo[1], J. M. Chow[1], J. M. Gambetta[2], Lev S. Bishop[1], B. R. Johnson[1], D. I. Schuster[1], J. Majer[3], A. Blais[4], L. Frunzio[1], S. M. Girvin[1] & R. J. Schoelkopf[1]

Quantum computers, which harness the superposition and entanglement of physical states, could outperform their classical counterparts in solving problems with technological impact—such as factoring large numbers and searching databases[1,2]. A quantum processor executes algorithms by applying a programmable sequence of gates to an initialized register of qubits, which coherently evolves into a final state containing the result of the computation. Building a quantum processor is challenging because of the need to meet simultaneously requirements that are in conflict: state preparation, long coherence times, universal gate operations and qubit readout. Processors based on a few qubits have been demonstrated using nuclear magnetic resonance[3–5], cold ion trap[6,7] and optical[8] systems, but a solid-state realization has remained an outstanding challenge. Here we demonstrate a two-qubit superconducting processor and the implementation of the Grover search and Deutsch–Jozsa quantum algorithms[1,2]. We use a two-qubit interaction, tunable in strength by two orders of magnitude on nanosecond timescales, which is mediated by a cavity bus in a circuit quantum electrodynamics architecture[9,10]. This interaction allows the generation of highly entangled states with concurrence up to 94 per cent. Although this processor constitutes an important step in quantum computing with integrated circuits, continuing efforts to increase qubit coherence times, gate performance and register size will be required to fulfil the promise of a scalable technology.

Over the past decade, superconducting circuits[11] have made considerable progress in all the requirements necessary for an electrically controlled, solid-state quantum computer. Coherence times[11,12] have risen by three orders of magnitude to ~1 μs, single-qubit gates[13,14] have reached error rates of 1%, engineered interactions[15–17] have produced two-qubit entanglement at a level of 60% concurrence[18], and qubit readout[18–20] has attained measurement fidelities of ~90%. However, combining these achievements in a single device remains challenging. One approach to integration is the quantum bus architecture[9,21,22], which uses a transmission line cavity to couple, control and measure qubits. We augment the architecture described in ref. 22 with flux-bias lines that tune individual qubit frequencies, permitting single-qubit phase gates. By pulsing the qubit frequencies to an avoided crossing where a $\sigma_z \otimes \sigma_z$ interaction turns on (σ_z is the Pauli z-operator), we are able to realize a two-qubit conditional phase (C-Phase) gate. Operation in the strong-dispersive regime[23] of circuit quantum electrodynamics (cQED) allows joint readout[24] that efficiently detects two-qubit correlations. Combined with single-qubit rotations, this enables tomography of the two-qubit state. Through improved understanding of spontaneous emission[25] and careful microwave engineering, we now attain state-of-the-art ~1 μs coherence times in a two-qubit device. This allows sufficient time to concatenate ~10 gates, realizing simple algorithms with fidelity greater than 80%.

Our processor, shown in Fig. 1a, is a four-port superconducting device comprising two transmon qubits[12,26] (Q_L and Q_R) inside a microwave cavity bus, and flux-bias lines proximal to each qubit. The cavity, normally off-resonance with the qubit transition frequencies f_L and f_R, couples the qubits by virtual photon exchange and shields them from the electromagnetic continuum. As previously demonstrated[22], microwave pulses resonant with f_L or f_R applied to the cavity input port provide frequency-multiplexed single-qubit x- and y-rotations with high fidelity[14] and selectivity. Pulsed measurement of the homodyne voltage V_H on the cavity output port provides qubit readout. The remaining two ports create local magnetic fields that tune the qubit transition frequencies. Each qubit has a split Josephson junction, so its frequency f depends on the flux Φ through the loop according to $hf \approx \sqrt{8E_C E_J^{max}|\cos(\pi\Phi/\Phi_0)|} - E_C$, where E_C is the charging energy, E_J^{max} is the maximum Josephson energy, h is Planck's constant, and Φ_0 is the flux quantum. By using short-circuited transmission lines with a bandwidth from d.c. to 2 GHz, we can tune f_L and f_R by many GHz using room temperature voltages V_L and V_R. Static tuning of qubit transitions using the flux-bias lines is demonstrated in Fig. 1b.

The spectrum of single excitations (Fig. 1b) shows the essential features of the cavity-coupled two-qubit Hamiltonian and allows determination of the relevant system parameters (see Methods). When the qubits are tuned to their maximum frequencies, point I, they are far detuned from the cavity and from each other, so that interactions are small. This point is therefore used for state preparation, single-qubit rotations and measurement, in the computational basis $|0, 0\rangle$, $|0, 1\rangle$, $|1, 0\rangle$ and $|1, 1\rangle$, where $|l, r\rangle$ denotes excitation level l (r) for Q_L (Q_R). Operation at this point is also desirable because it is a flux 'sweet spot'[12] for both qubits, providing long coherence, with relaxation and dephasing times $T_{1,L(R)} = 1.3(0.8)$ μs and $T^*_{2,L(R)} = 1.8(1.2)$ μs, respectively. Tuning Q_R into resonance with the cavity, point IV in Fig. 1b, reveals a vacuum Rabi splitting[10] from which the qubit–cavity interaction strength is extracted. Tuning Q_R into resonance with Q_L, point III, shows an avoided crossing resulting from a cavity-mediated, qubit–qubit transverse interaction[9,27] investigated previously[22]. In this work, we perform two-qubit gates at point II, where no interactions are immediately apparent on examining the one-excitation manifold.

However, a useful two-qubit interaction is revealed in the two-excitation spectrum, Fig. 2a. As V_R is swept away from point I, the non-computational higher-level transmon excitation $|0, 2\rangle$ decreases more rapidly than the computational state $|1, 1\rangle$, and these states would become degenerate at point II. But as shown in Fig. 2b, there is a large (160 MHz) cavity-mediated interaction between these

[1]Departments of Physics and Applied Physics, Yale University, New Haven, Connecticut 06511, USA. [2]Department of Physics and Astronomy and Institute for Quantum Computing, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. [3]Atominstitut der Österreichischen Universitäten, TU-Wien, A-1020 Vienna, Austria. [4]Département de Physique, Université de Sherbrooke, Sherbrooke, Québec J1K 2R1, Canada.

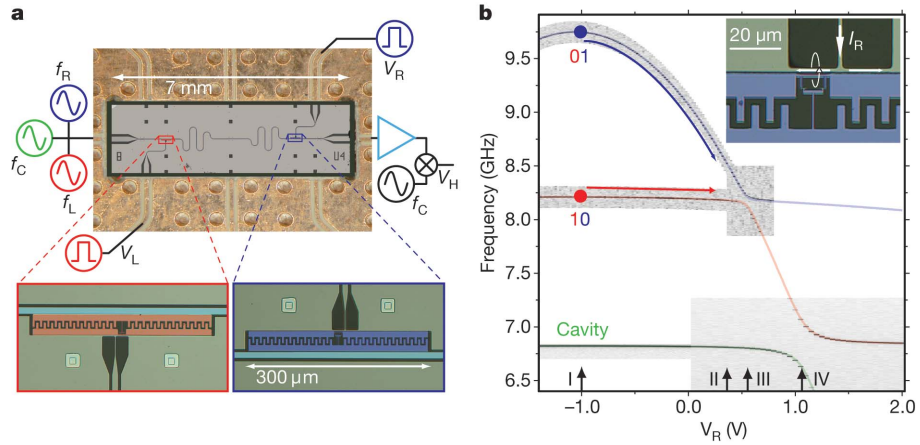


Figure 1 | Two-qubit cQED device, and cavity/qubit characterization. **a**, Optical micrograph of four-port device with a coplanar waveguide cavity bus coupling transmon qubits Q_L and Q_R (coloured red and blue in insets), and local flux-bias lines providing fast qubit tuning. Microwave pulses at the qubit transition frequencies f_L and f_R drive single-qubit rotations, and a pulsed measurement of the cavity homodyne voltage V_H (at frequency f_C) provides two-qubit readout. The flux-bias lines (bottom-left and top-right ports) are coplanar waveguides with short-circuit termination next to their target qubit. The termination geometry allows currents (I_L and I_R) on the lines to couple flux through the split junctions (**b**, inset). **b**, Grey-scale images of cavity transmission and of qubit spectroscopy as a function of V_R, showing local tuning of Q_R across the avoided crossing with Q_L (point III) and across the vacuum Rabi splitting with the cavity (point IV). Semi-transparent lines are theoretical best fits obtained from numerical diagonalization of a generalized Tavis–Cummings Hamiltonian[28]. Points I and II are the operating points of the processor. Preparation, single-qubit operations and measurements are performed at point I, and a C-Phase gate is achieved by pulsing into point II. Numerals indicate excitation level of Q_L (red) and Q_R (blue) in the spectroscopy at point I.

levels, producing a frequency shift $\zeta/2\pi$ of the lower branch with respect to the sum $f_L + f_R$, in good agreement with a numerical diagonalization of the generalized Tavis–Cummings Hamiltonian[28] (see Methods).

This shift is the mechanism of our conditional phase gate. Flux pulses, adiabatic with respect to the $|1, 1\rangle \leftrightarrow |0, 2\rangle$ avoided crossing, produce phase gates

$$U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{i\phi_{01}} & 0 & 0 \\ 0 & 0 & e^{i\phi_{10}} & 0 \\ 0 & 0 & 0 & e^{i\phi_{11}} \end{pmatrix}$$

in the computational Hilbert space. Here, $\phi_{lr} = 2\pi\int\delta f_{lr}(t)\mathrm{d}t$ is the dynamical phase acquired by $|l, r\rangle$, and δf_{lr} is the deviation of f_{lr} from its value at point I. A V_R pulse into point II such that $\int\zeta(t)\mathrm{d}t = (2n+1)\pi$ with integer n implements a C-Phase gate, because $\phi_{11} = \phi_{01} + \phi_{10} - \int\zeta(t)\mathrm{d}t$. This method of realizing a C-Phase gate by adiabatically using the avoided crossing between computational and non-computational states is generally applicable to qubit implementations with finite anharmonicity, such as transmons[12] or phase qubits[13]. A similar approach involving higher excitation levels but with non-adiabatic pulses was previously proposed[29]. The negative anharmonicity permits the phase gate at point II to occur before the onset of transverse coupling at point III.

Control of ζ by two orders of magnitude provides an excellent on-off ratio for the C-Phase gate. Measurements of ζ obtained from spectroscopy and from time-domain experiments show very good agreement (Fig. 2c). The time-domain method measures the difference in the precession frequency of Q_L in two Ramsey-style experiments, where a V_R-pulse of varying duration (0–100 ns) is inserted between $\pi/2$ rotations of Q_L, with Q_R either in the ground state $|0\rangle$ or excited into state $|1\rangle$. Using the time-domain approach, we measure a residual $\zeta/2\pi \approx 1.2$ MHz at point I (star in Fig. 2c). The theoretical ζ obtained by numerical diagonalization shows reasonable agreement with the data, except for a scale factor that is probably due to higher modes of the cavity[25], not included in the calculation.

The controlled phase interaction allows universal two-qubit gates. As an example, we produce entangled states on demand (Fig. 3). The pulse sequence in Fig. 3a generates any of the four Bell states,

$$|\Psi^{\pm}\rangle = \tfrac{1}{\sqrt{2}}(|0, 0\rangle \pm |1, 1\rangle) \quad |\Phi^{\pm}\rangle = \tfrac{1}{\sqrt{2}}(|0, 1\rangle \pm |1, 0\rangle)$$

depending on the choice of C-Phase gate $\mathrm{c}U_{ij}$ applied $(\mathrm{c}U_{ij}|l, r\rangle = (-1)^{\delta_{il}\delta_{jr}}|l, r\rangle$, where δ is Kronecker's delta). We achieve $\int\zeta(t)\mathrm{d}t = \pi$ by tuning the amplitude of a 30 ns V_R-pulse close to point II and back. During the pulse, Q_R acquires a large dynamical phase $\phi_{01} \approx -60\pi$. The four $\mathrm{c}U_{ij}$ gates differ by whether ϕ_{01} and ϕ_{10} are even or odd multiples of π. We tune ϕ_{01} over a 2π range by adjusting the rising and falling edges of the pulse, and ϕ_{10} by varying the amplitude of a simultaneous weak V_L-pulse (Supplementary Fig. 3). The conditional phase $\int\zeta(t)\mathrm{d}t$ is largely independent of these two adjustments.

To detect the entanglement, we reconstruct the two-qubit density matrix ρ by quantum state tomography using joint dispersive readout[9,22,24]. A pulsed measurement of the homodyne voltage V_H measures the operator:

$$M = \beta_1\sigma_z^L + \beta_2\sigma_z^R + \beta_{12}\sigma_z^L \otimes \sigma_z^R$$

Operation in the strong-dispersive regime[23,24] makes the three constant coefficients have approximately the same magnitude, $|\beta_{12}| \approx |\beta_1|, |\beta_2|$, enhancing sensitivity to two-qubit correlations. A complete set of 15 linearly independent operators is built using single-qubit rotations before measuring M. An ensemble average of each operator is obtained by executing the sequence in Fig. 3a 450,000 times. The 15 average values are then input to a maximum-likelihood estimator of ρ (Supplementary Information).

The inferred density matrices ρ_{ml} reveal in all four cases (Fig. 3b–e) a high degree of two-qubit entanglement, which we quantify using concurrence[30], C. Values are listed in Fig. 3 legend, along with the metrics of purity $P(\rho) = \mathrm{Tr}(\rho^2)$ and fidelity to the target state $|\psi\rangle$, $F(\rho, \psi) = \langle\psi|\rho|\psi\rangle$. Note that there are several common definitions of fidelity in the literature, and our definition is the square of the fidelity used in refs 18 and 24. The quoted values significantly extend the state of the art for solid-state entanglement[18], and provide evidence that we have a high-fidelity universal set of two-qubit gates.

One- and two-qubit gates can be concatenated to realize simple algorithms, such as Grover's quantum search[1,2] shown in Fig. 4. Given a function $f(x)$ on the two-bit set $x \in \{00, 01, 10, 11\}$ such that $f(x) = 0$ except at some x_0, where $f(x_0) = 1$, this well-known algorithm can determine x_0 with a single call of an oracle O that encodes

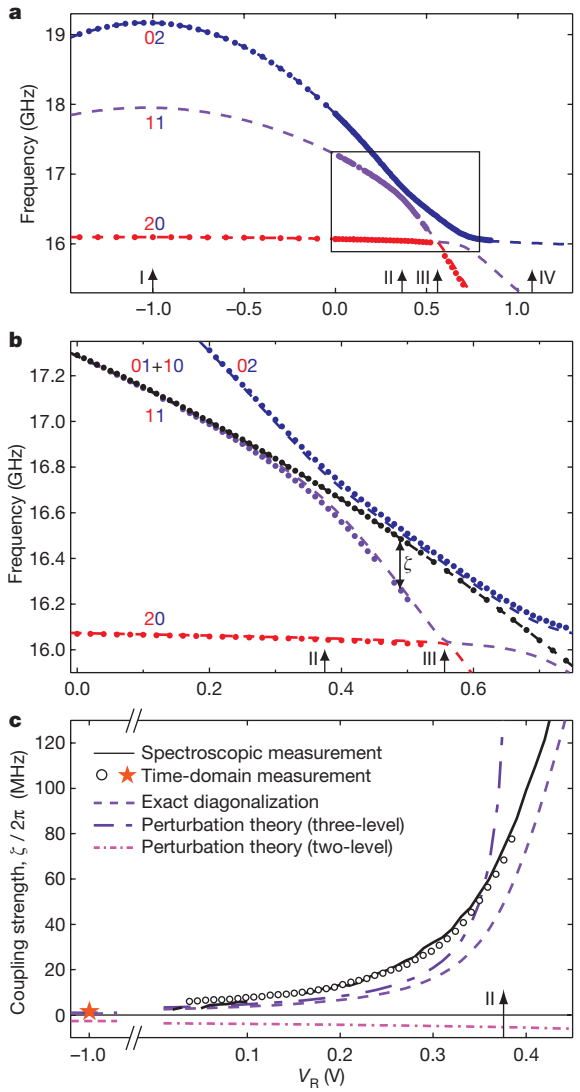


Figure 2 | Origin and characterization of the controlled-phase gate. a, Flux dependence of transition frequencies from the ground state $|0, 0\rangle$ to the two-excitation manifold. Red (blue) numerals indicate the excitation level of the left (right) transmon for each transition. Two-tone spectroscopy measurements[12] (points) show an avoided crossing between the computational state $|1, 1\rangle$ and the non-computational state $|0, 2\rangle$ at point II, in good agreement with numerical diagonalization of the Hamiltonian (dashed curves). **b**, This avoided crossing causes the transition frequency to $|1, 1\rangle$ to deviate from the sum of the transition frequencies to $|0, 1\rangle$ and $|1, 0\rangle$. **c**, The coupling strength $\zeta/2\pi = f_{01} + f_{10} - f_{11}$ of the effective $\sigma_z^L \otimes \sigma_z^R$ interaction, obtained both from spectroscopy (solid curve) and from time-domain experiments (points; see text for details). Numerical diagonalization and perturbation theory (Supplementary Information) for three-level transmons agree reasonably with data. The perturbation calculation diverges at the avoided crossing. Perturbation theory for two-level qubits gives the wrong magnitude and sign for ζ, and demonstrates that the higher transmon excitations are necessary for the interaction. Time-domain measurement and theory both give $\zeta/2\pi \approx 1.2$ MHz at point I. The tunability of ζ over two orders of magnitude provides an excellent on-off ratio for the two-qubit C-Phase gate.

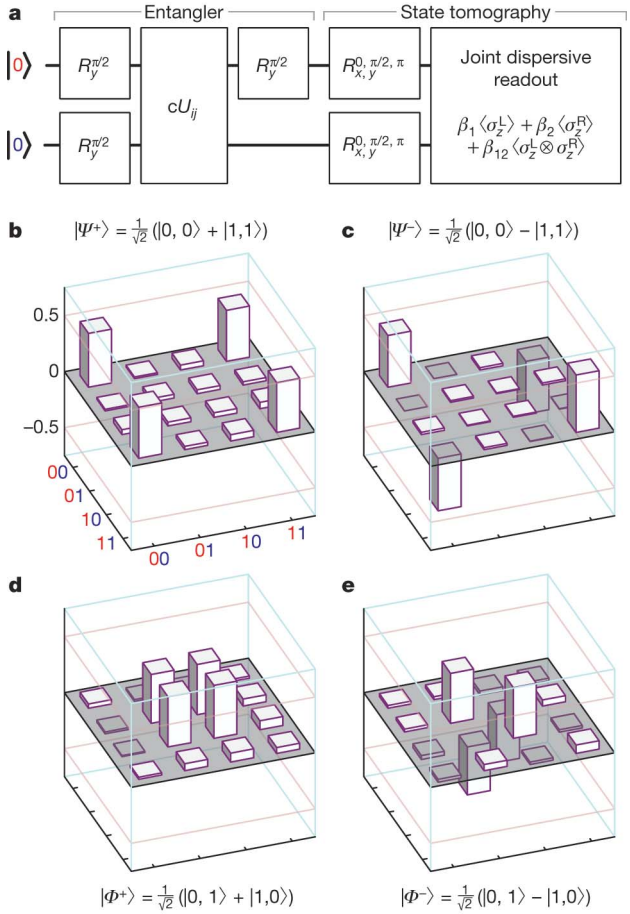


Figure 3 | Entanglement on demand. a, Gate sequence generating two-qubit entanglement and detection via quantum state tomography. Starting from $|0, 0\rangle$, simultaneous $\pi/2$ rotations on both qubits create an equal superposition of the four computational states. A C-Phase cU_{ij} then phase shifts $|i, j\rangle$ in the superposition and produces entanglement. A final $\pi/2$ rotation on Q_L evolves the entangled state into one of the four Bell states depending on the cU_{ij} applied. **b–e**, Real part of maximum-likelihood density matrix ρ_{ml} of the entangler output for cU_{10}, cU_{00}, cU_{11} and cU_{01}, respectively (imaginary elements of ρ_{ml} are less than 0.03, 0.02, 0.07, 0.08). Extracted metrics for the four entangler outputs include concurrence $C = 0.88 \pm 0.02$, 0.94 ± 0.01, 0.86 ± 0.02, 0.81 ± 0.04, purity $P = 0.87 \pm 0.02$, 0.92 ± 0.02, 0.88 ± 0.02, 0.79 ± 0.03, and fidelity to the ideal Bell state $F = 0.91 \pm 0.01$, 0.94 ± 0.01, 0.90 ± 0.01, 0.87 ± 0.02. The uncertainties correspond to the standard deviation in 16 repetitions of generation-tomography for each entangler.

$f(x)$ in a quantum phase, $O|x\rangle = (-1)^{f(x)}|x\rangle$. The oracle for $x_0 = ij$ is the C-Phase gate cU_{ij}.

We can examine the functioning of the algorithm by interrupting it after each step and performing state tomography. Figure 4b–g shows all the features of a quantum processor, namely the use of maximally superposed states to exploit quantum parallelism (Fig. 4c), the encoding of information in the entanglement between qubits (Fig. 4d, e), and the interference producing an answer represented in a final computational basis state. The fidelity of the final state (Fig. 4g) to the expected output ($|1, 0\rangle$ for the case $O = cU_{10}$ shown) is 85%. Similar performance is obtained for the other three oracles (Table 1).

We have also implemented the Deutsch–Jozsa algorithm[1,2]. The two-qubit version of this algorithm determines whether an unknown function $f_i(x)$, mapping a one-bit input to a one-bit output, is constant ($f_0(x) = 0$ or $f_1(x) = 1$) or balanced ($f_2(x) = x$ or $f_3(x) = 1 - x$) with a single call of the function. The algorithm applies the function once to a superposition of the two possible inputs and uses quantum phase kick-back[2] to encode the result in the final state of one qubit (Q_L) while leaving the other untouched (Q_R). The gate sequence realizing the algorithm and the output tomographs for the four cases are shown in Supplementary Fig. 1.

The performance of both algorithms is summarized in Table 1. Although there are undoubtedly significant systematic errors remaining, the overall fidelity is similar to that expected from the ratio (~100 ns/1 μs) of the total duration of gate sequences to the qubit coherence times. The detailed error budget will be addressed in future work using quantum process tomography.

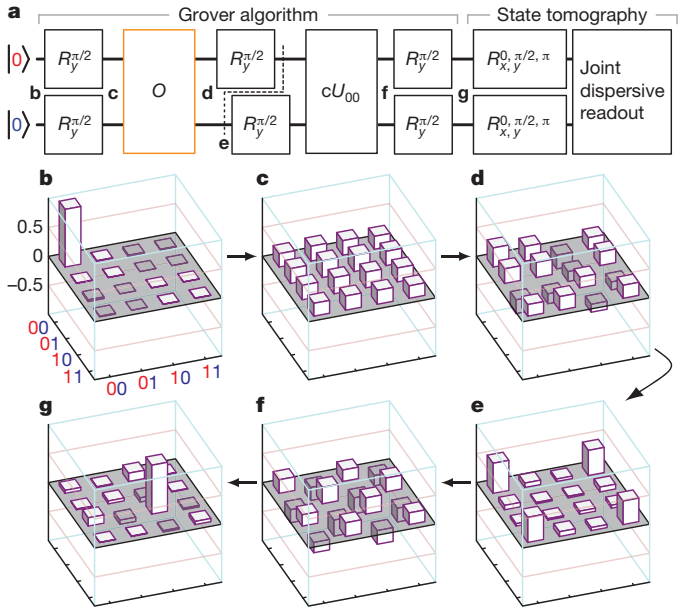


Figure 4 | Implementation of Grover's search algorithm. **a**, Concatenation of single-qubit and C-Phase gates implementing one iteration of Grover searching. Without loss of generality, we have replaced the Walsh–Hadamard transformations $W = R_x^{\pi} R_y^{\pi/2}$ in the usual description of the algorithm[1,2] with $R_y^{\pi/2}$ rotations in order to eliminate six single-qubit rotations and complete the sequence in 104 ns. (Supplementary Fig. 3 shows the microwave and flux pulses implementing the sequence.) The orange box is the oracle $O = cU_{ij}$ that encodes the solution $x_0 = ij$ to the search problem in a quantum phase. Note that the first half of the algorithm is identical to the entangling sequence in Fig. 3, while the second half is essentially its mirror image. **b–g**, Real part of ρ_{ml} obtained by state tomography after each step of the algorithm with oracle $O = cU_{10}$. Starting from $|0, 0\rangle$ (**b**), the qubits are simultaneously rotated into a maximal superposition state (**c**). The oracle then marks the solution, $|1, 0\rangle$, by inverting its phase (**d**). The $R_y^{\pi/2}$ rotation on Q_L turns the state into the Bell state $|\Psi^+\rangle$, demonstrating that the state is highly entangled at this stage. The $R_y^{\pi/2}$ rotation on Q_R produces a state identical to (**d**) (data not shown). The application of cU_{00} undoes the entanglement, producing a maximal superposition state (**f**). The final rotations yield an output state (**g**) with fidelity $F = 85\%$ to the correct answer, $|1, 0\rangle$.

In summary, we have demonstrated two-qubit quantum algorithms using a superconducting circuit. The incorporation of local flux control and joint-dispersive readout into cQED, together with a tenfold increase in qubit coherence over previous two-qubit devices, has enabled on-demand generation and detection of entanglement and the implementation of the Grover and Deutsch–Jozsa algorithms. The present architecture can be immediately expanded to several qubits with controllable $\sigma_z \otimes \sigma_z$ interactions between nearest-frequency neighbours, placing within reach the generation of Greenberger–Horne–Zeilinger states and the exploration of basic concepts of quantum error correction[1,2].

METHODS SUMMARY

Device fabrication. A 180 nm Nb film was d.c.-magnetron sputtered on the epi-polished surface of an R-plane corundum (α-Al_2O_3) wafer (2 inches diameter, 430 μm thickness). Coplanar waveguide structures (cavity and flux-bias lines) were patterned by optical lithography and fluorine-based reactive ion etching of Nb. Transmon features (interdigitated capacitors and split junctions) were patterned on 2 mm × 7 mm chips using electron-beam lithography, double-angle evaporation of Al (20/90 nm) with intermediate oxidation (15% O_2 in Ar at 15 torr for 12 min), and lift-off.

A completed device was cooled to 13 mK in a ^{3}He-^{4}He dilution refrigerator. The refrigerator wiring is shown in Supplementary Fig. 2. Careful microwave engineering of the sample holder and on-chip wirebonding across ground planes were required to suppress spurious resonance modes on- and off-chip. Simulations using Sonnet software assisted this iterative process. The sample was enclosed in two layers of Cryoperm magnetic shielding, allowing high-fidelity operation of the processor during unattended overnight runs.

cQED theory. The Tavis–Cummings[28] Hamiltonian generalized to multi-level transmon qubits[26] is:

$$H = \omega_C a^\dagger a + \sum_{q\in\{L,R\}} \left(\sum_{j=0}^{N} \omega_{0j}^q |j\rangle_q\langle j|_q + (a + a^\dagger) \sum_{j,k=0}^{N} g_{jk}^q |j\rangle_q\langle k|_q \right) \quad (1)$$

Here, ω_C is the bare cavity frequency, $\omega_{0j}^q = \omega_{0j}(E_{Cq}, E_{Jq})$ is the transition frequency for qubit q from ground to excited state j, and $g_{jk}^q = g_q n_{jk}(E_{Cq}, E_{Jq})$, with g_q a bare qubit–cavity coupling and n_{jk} a level-dependent coupling matrix element. The dependence of these parameters on qubit charging energy E_{Cq} and Josephson energy E_{Jq} is indicated. The flux control enters through $E_{Jq} = E_{Jq}^{max}|\cos(\pi\Phi_q/\Phi_0)|$, with Φ_q the flux through the qubit loop, and a linear flux–voltage relation $\Phi_q = \alpha_{qL} V_L + \alpha_{qR} V_R + \Phi_{q,0}$, accounting for crosstalk and offsets. (Crosstalk, ~30%, probably results from spatial distribution of flux-bias return currents on the ground plane.) The above parameters are tightly constrained by the spectroscopy and transmission data shown (Figs 1b, 2a and b) and transmission data (not shown) for the Q_L-cavity vacuum Rabi splitting. Simultaneously fitting the spectra given by numerical diagonalization of the Hamiltonian (truncated to $N = 5$ qubit levels and 5 cavity photons) to these data gives $E_{JL(R)}^{max}/h = 28.48(42.34)$ GHz, $E_{CL(R)}/h = 317(297)$ MHz, $g_{L(R)}/2\pi = 199(183)$ MHz. Cavity parameters are $\omega_C/2\pi = 6.902$ GHz and linewidth $\kappa/2\pi = 1$ MHz.

Received 11 March; accepted 5 May 2009.
Published online 28 June.

1. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, 2000).
2. Kaye, P., Laflamme, R. & Mosca, M. *An Introduction to Quantum Computing* (Oxford Univ. Press, 2007).
3. Chuang, I. L., Vandersypen, L. M. K., Zhou, X., Leung, D. W. & Lloyd, S. Experimental realization of a quantum algorithm. *Nature* **393**, 143–146 (1998).
4. Jones, J. A., Mosca, M. & Hansen, R. H. Implementation of a quantum search algorithm on a quantum computer. *Nature* **393**, 344–346 (1998).
5. Chuang, I. L., Gershenfeld, N. & Kubinec, M. Experimental implementation of fast quantum searching. *Phys. Rev. Lett.* **80**, 3408–3411 (1998).
6. Guide, S. *et al.* Implementation of the Deutsch-Jozsa algorithm on an ion-trap quantum computer. *Nature* **421**, 48–50 (2003).
7. Brickman, K.-A. *et al.* Implementation of Grover's quantum search algorithm in a scalable system. *Phys. Rev. A* **72**, 050306(R) (2005).

Table 1 | Summary of algorithmic performance

Element		Grover search oracle*				Deutsch–Jozsa function†			
		f_{00}	f_{01}	f_{10}	f_{11}	f_0	f_1	f_2	f_3
$\langle 0,0\|\rho\|0,0\rangle$	Ideal	1	0	0	0	0	0	1	1
	Measured	0.81(1)	0.08(1)	0.07(2)	0.065(7)	0.010(3)	0.014(5)	0.909(6)	0.841(9)
$\langle 0,1\|\rho\|0,1\rangle$	Ideal	0	1	0	0	0	0	0	0
	Measured	0.066(7)	0.802(9)	0.05(1)	0.054(8)	0.012(4)	0.008(4)	0.031(8)	0.04(2)
$\langle 1,0\|\rho\|1,0\rangle$	Ideal	0	0	1	0	1	1	0	0
	Measured	0.08(1)	0.05(1)	0.82(2)	0.07(1)	0.93(1)	0.93(1)	0.05(1)	0.04(1)
$\langle 1,1\|\rho\|1,1\rangle$	Ideal	0	0	0	1	0	0	0	0
	Measured	0.05(2)	0.07(1)	0.06(1)	0.81(1)	0.05(1)	0.04(1)	0.012(9)	0.07(2)

Fidelity of the reconstructed output states of the Grover and Deutsch–Jozsa algorithms to their ideal outputs. These results suggest that, if combined with single-shot readout, the two algorithms executed with this processor would give the correct answer with probability far exceeding the 50% success probability of the best classical algorithms limited to single calls of the oracle[7] or function.
* Uncertainties are based on 10 repetitions.
† Uncertainties are based on 8 repetitions.

8. Kwiat, P. G., Mitchell, J. R., Schwindt, P. D. D. & White, A. G. Grover's search algorithm: an optical approach. *J. Mod. Opt.* **47,** 257–266 (2000).
9. Blais, A., Huang, R.-S., Wallraff, A., Girvin, S. M. & Schoelkopf, R. J. Cavity quantum electrodynamics for superconducting electrical circuits: an architecture for quantum computation. *Phys. Rev. A* **69,** 062320 (2004).
10. Wallraff, A. *et al.* Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics. *Nature* **431,** 162–167 (2004).
11. Clarke, J. & Wilhelm, F. K. Superconducting quantum bits. *Nature* **453,** 1031–1042 (2008).
12. Schreier, J. A. *et al.* Suppressing charge noise decoherence in superconducting charge qubits. *Phys. Rev. B* **77,** 180502(R) (2008).
13. Lucero, E. *et al.* High-fidelity gates in a single Josephson qubit. *Phys. Rev. Lett.* **100,** 247001 (2008).
14. Chow, J. M. *et al.* Randomized benchmarking and process tomography for gate errors in a solid-state qubit. *Phys. Rev. Lett.* **102,** 090502 (2009).
15. Yamamoto, T., Pashkin, Yu. A., Astafiev, O., Nakamura, Y. & Tsai, J. S. Demonstration of conditional gate operation using superconducting charge qubits. *Nature* **425,** 941–944 (2003).
16. Plantenberg, J. H., de Groot, P. C., Harmans, C. J. P. M. & Mooij, J. E. Demonstration of controlled-NOT quantum gates on a pair of superconducting quantum bits. *Nature* **447,** 836–839 (2007).
17. Niskanen, A. O. *et al.* Quantum coherent tunable coupling of superconducting qubits. *Science* **316,** 723–726 (2007).
18. Steffen, M. *et al.* Measurement of the entanglement of two superconducting qubits via state tomography. *Science* **313,** 1423–1425 (2006).
19. Siddiqi, I. *et al.* RF-driven Josephson bifurcation amplifier for quantum measurement. *Phys. Rev. Lett.* **93,** 207002 (2004).
20. McDermott, R. *et al.* Simultaneous state measurement of coupled Josephson phase qubits. *Science* **307,** 1299–1302 (2005).
21. Sillanpää, M. A., Park, J. I. & Simmonds, R. W. Coherent quantum state storage and transfer between two phase qubits via a resonant cavity. *Nature* **449,** 438–442 (2007).
22. Majer, J. *et al.* Coupling superconducting qubits via a cavity bus. *Nature* **449,** 443–447 (2007).
23. Schuster, D. I. *et al.* Resolving photon number states in a superconducting circuit. *Nature* **445,** 515–518 (2007).
24. Filipp, S. *et al.* Two-qubit state tomography using a joint dispersive read-out. *Phys. Rev. Lett.* **102,** 200402 (2009).
25. Houck, A. A. *et al.* Controlling the spontaneous emission of a superconducting transmon qubit. *Phys. Rev. Lett.* **101,** 080502 (2008).
26. Koch, J. *et al.* Charge-insensitive qubit design derived from the Cooper pair box. *Phys. Rev. A* **76,** 042319 (2007).
27. Blais, A. *et al.* Quantum-information processing with circuit quantum electrodynamics. *Phys. Rev. A* **75,** 032329 (2007).
28. Tavis, M. & Cummings, F. W. Exact solution for an *N*-molecule-radiation-field Hamiltonian. *Phys. Rev.* **170,** 379–384 (1968).
29. Strauch, F. W. *et al.* Quantum logic gates for coupled superconducting phase qubits. *Phys. Rev. Lett.* **91,** 167005 (2003).
30. Wootters, W. K. Entanglement of formation of an arbitrary state of two qubits. *Phys. Rev. Lett.* **80,** 2245–2248 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank V. Manucharyan and E. Boaknin for experimental contributions, and M. H. Devoret, I. L. Chuang and A. Nunnenkamp for discussions. This work was supported by LPS/NSA under ARO contract W911NF-05-1-0365, and by the NSF under grants DMR-0653377 and DMR-0603369. We acknowledge additional support from CIFAR, MRI, MITACS and NSERC (J.M.G.), NSERC, CIFAR and the Alfred P. Sloan Foundation (A.B.), and from CNR-Istituto di Cibernetica, Pozzuoli, Italy (L.F.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.J.S. (robert.schoelkopf@yale.edu).

LETTERS

Direct observation of correlations between individual photon emission events of a microcavity laser

J. Wiersig[1]†, C. Gies[1], F. Jahnke[1], M. Aßmann[2], T. Berstermann[2], M. Bayer[2], C. Kistner[3], S. Reitzenstein[3], C. Schneider[3], S. Höfling[3], A. Forchel[3], C. Kruse[1], J. Kalden[1] & D. Hommel[1]

Lasers are recognized for coherent light emission, the onset of which is reflected in a change in the photon statistics[1]. For many years, attempts have been made to directly measure correlations in the individual photon emission events of semiconductor lasers[2,3]. Previously, the temporal decay of these correlations below or at the lasing threshold was considerably faster than could be measured with the time resolution provided by the Hanbury Brown/Twiss measurement set-up[4] used. Here we demonstrate a measurement technique using a streak camera that overcomes this limitation and provides a record of the arrival times of individual photons. This allows us to investigate the dynamical evolution of correlations between the individual photon emission events. We apply our studies to micropillar lasers[5] with semiconductor quantum dots[2,3,6–8] as the active material, operating in the regime of cavity quantum electrodynamics[9]. For laser resonators with a low cavity quality factor, *Q*, a smooth transition from photon bunching to uncorrelated emission with increasing pumping is observed; for high-*Q* resonators, we see a non-monotonic dependence around the threshold where quantum light emission can occur. We identify regimes of dynamical anti-bunching of photons in agreement with the predictions of a microscopic theory that includes semiconductor-specific effects.

Conventional lasers (like gas lasers or semiconductor edge emitters) have a well-defined laser threshold that manifests itself in a sharp increase in the number of emitted photons. Also, the classical and quantum mechanical properties of the emitted light undergo pronounced changes at the laser threshold. The coherence time of the radiation increases strongly and the photon statistics reflects a transition from thermal radiation to coherent emission. The origin of the laser threshold is the loss of the majority of spontaneously emitted photons into non-lasing modes.

Cavity quantum electrodynamics (QED) has opened the door to a new class of lasers that are based on the altered spontaneous-emission properties of the active material, these alterations being due to the modified electromagnetic environment[9]. The Purcell effect makes it possible to inhibit the spontaneous emission into non-lasing modes and enhance that into lasing modes. As a result, in current state-of-the-art semiconductor microcavity lasers[2,3,5–8], the kink in the output intensity as a function of pump rate—previously taken to be a clear indication of a laser threshold—is strongly diminished or even lost.

There are fundamental reasons[10] why the laser threshold is no longer well defined when cavity QED effects become important and the spontaneous emission into non-lasing modes is strongly suppressed. In this regime, the number of photons in the cavity and the number of electronic excitations in the active material are no longer much larger than one, and quantum fluctuations have a central role. Recent advances in the technical realization of semiconductor microcavity lasers have exposed an urgent need for new tools with which to characterize these devices, which operate in a regime that differs fundamentally from that of conventional lasers. This refers both to experimental techniques for characterizing photon correlations and their dynamical evolution on the relevant timescales in semiconductor systems and to quantum statistical models that include semiconductor-specific effects.

The statistical properties of a light beam can be characterized using the second-order, or intensity, autocorrelation function. A classical version of this correlation function was first used by Hanbury Brown and Twiss[4]. Shortly after, Glauber[1] introduced a quantum mechanical description of the second-order correlation function

$$G^{(2)}(t,\tau)=\langle b^{\dagger}(t)b^{\dagger}(t+\tau)b(t+\tau)b(t)\rangle \quad (1)$$

which is often used in its normalized form

$$g^{(2)}(t,\tau)=\frac{G^{(2)}(t,\tau)}{\langle b^{\dagger}(t)b(t)\rangle\langle b^{\dagger}(t+\tau)b(t+\tau)\rangle} \quad (2)$$

Here $b^{\dagger}$ and b are respectively photon creation and annihilation operators of the optical mode of interest, and angle brackets denote the quantum mechanical expectation value with respect to the density operator, ρ.

Under stationary conditions (no t dependence in equations (1) and (2)), the second-order correlation function at zero delay, $\tau = 0$, directly reflects the statistical properties of the emitted light. For conventional lasers with small β factor (describing the fraction of the spontaneous emission into the laser mode relative to the total spontaneous emission), a sharp transition from thermal, $g^{(2)}(0) = 2$, to coherent light emission, $g^{(2)}(0) = 1$, is expected when passing the threshold to lasing. For thermal light, it is also known that $g^{(2)}(\tau)$, as a function of the delay time, decays to unity on the timescale of the coherence time[11]. A drawback of previous investigations of photon correlations in the laser emission was the finite time resolution of the experiments. In a standard Hanbury Brown/Twiss measurement with avalanche photodiodes, the time resolution is about 500 ps, whereas the coherence time of semiconductor lasers below threshold is less than 10 ps (ref. 3). In this example, the experiment averages over a time interval, in which much of the decay of $g^{(2)}(\tau)$ from two to one has taken place. This leads to an underestimation of $g^{(2)}(0)$ and explains the observed below-threshold reduction of the correlation function in previous experiments[2,3].

We consider single-photon sources an important, albeit different, topic because photon anti-bunching is demonstrated in the spontaneous-emission regime of single-atom[12] or single-quantum-dot[13] systems (although the improved time resolution of our new measurement scheme will be beneficial when the Purcell effect is used to shorten the spontaneous lifetime). Earlier experiments on

[1]Physics Department, University of Bremen, 28334 Bremen, Germany. [2]Experimentelle Physik II, Technische Universität Dortmund, 44221 Dortmund, Germany. [3]Technische Physik, University of Würzburg, 97074 Würzburg, Germany. †Present address: Institute for Theoretical Physics, University of Magdeburg, 39016 Magdeburg, Germany.

intensity fluctuations in conventional semiconductor lasers[14] addressed noise in continuous-variable measurements and led to the seminal observation of squeezing as a true quantum effect. Corresponding theoretical investigations based on a Langevin theory were presented, for example, in ref. 15. In this Letter, however, we demonstrate correlations between single-photon events.

For this purpose, we introduce a new measurement scheme that reveals the arrival times of the individual photons in a light beam, thereby allowing us to reconstruct the second-order (and even higher-order) photon correlation functions. The capabilities of this method are demonstrated in the characterization of quantum-dot-based high-Q microcavity lasers. These systems can exhibit a strongly reduced kink in the input and output power traces, as well as the absence of a sharp transition in $g^{(2)}(0)$ between the values reflecting thermal and Poissonian photon statistics, making the threshold no longer well defined. In addition to measurements of $g^{(2)}(\tau = 0)$ without the limitations discussed above, we study the τ dependence of photon correlations and find an unexpected dynamical anti-bunching effect. To verify our results and to address the underlying physics, we use a microscopic theory for photon correlations in the emission of quantum-dot-based microcavity lasers. In the past, the use of single or few quantum-dot emitters for manipulating the photon emission statistics has been suggested, for example in refs 16, 17.

We obtain our experimental results by time-resolved single-photon counting with a resolution down to 2 ps for $g^{(2)}(\tau)$ measurements without the above-discussed limitations. After non-resonant pulsed excitation of the sample, which is held at 6 K, the subsequent emission is recorded with a streak camera in photon counting mode. A fast silicon charge-coupled-device camera is used to read out the streak camera screen. The temporal resolution is achieved by applying high voltage gradients to the horizontal and vertical camera capacitors, causing time-dependent deflections of the photoelectrons inside the streak camera. The camera is synchronized with the exciting laser, such that the sweep period is equal to the laser repetition rate.

Examples of streak camera pictures are shown in Fig. 1. Each dot in Fig. 1b (enlarged for better visibility) corresponds to a single photon detection. For the determination of the second-order correlation function, within each streak the number of photon pairs, the first detected at time t and the second delayed by a time τ, is counted. Normalization is obtained using the product of the average intensities at times t and $t+\tau$ according to equation (2). The resulting normalized correlation function, $g^{(2)}(t,\tau)$, is then averaged over a large number of streaks. Finally, $g^{(2)}(\tau)$ is obtained by averaging over all emission times t. The pictures in Fig. 1 were recorded for a low-Q cavity using III-V semiconductor material.

Extracted results for the second-order photon correlation function, $g^{(2)}(\tau = 0)$, are provided in Fig. 2, together with the corresponding input–output curves, for three different micropillars with cavity qualities Q that increase from Fig. 2a to Fig. 2c. A few characteristic features can be observed. Comparing the input–output curves, the jump from spontaneous to stimulated emission is quite small—about one order of magnitude for the II-VI cavity and two orders of magnitude for the III-V cavities. Furthermore, for the low-Q III-V cavity the jump is rather sharp, whereas for the other two cavities it is smoother. This indicates that the studied samples are strongly influenced by cavity QED effects. From the sample parameters, it may be expected that the low-Q III-V cavity shows the behaviour closest to a conventional laser, as exemplified by the relatively sharp jump in the input–output curve. This jump is already softened for the II-VI cavity, but the variation with cavity parameters suggests that the input–output curve does not provide a unique characterization of the light emission.

Much more insight can be obtained from the zero-delay second-order correlation function. For the II-VI cavity, it shows a transition from values slightly below two (characteristic for thermal light) to those approaching one (suggesting Poisson statistics) in the onset region of stimulated emission. For low excitation, $g^{(2)}(\tau = 0)$ saturates at a value of 1.95, which is less than two as expected for a

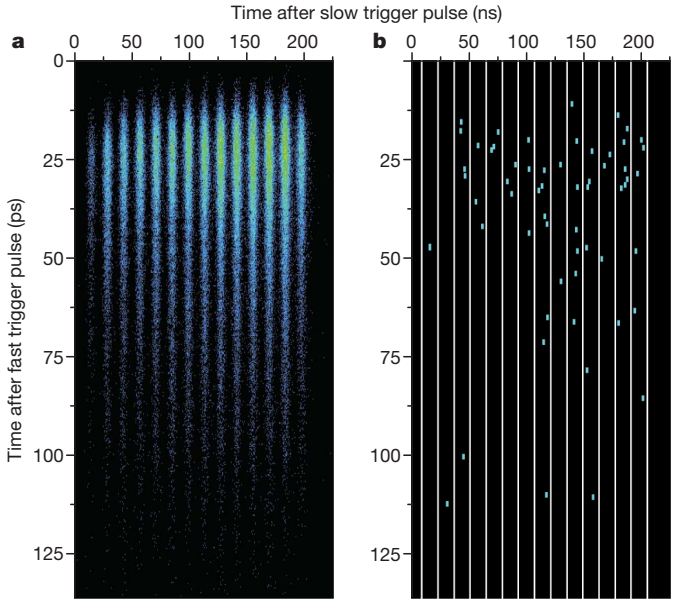


Figure 1 | Time-resolved photon counting statistics. **a**, **b**, Statistics integrated over 60,000 screens (**a**) and for a single screen of the streak camera (**b**), revealing temporal correlations between individual photons. The vertical axes show the temporal resolution after each excitation pulse. Along the horizontal axes the emission after subsequent excitation pulses is displaced. **a** visualizes the statistically averaged emission dynamics after pulsed excitation. For the determination of photon correlations, the signal is read out after only one screen has been filled with a set of streaks, as shown in **b** for 14 excitation pulses. Regions associated with different excitation pulses are separated by lines.

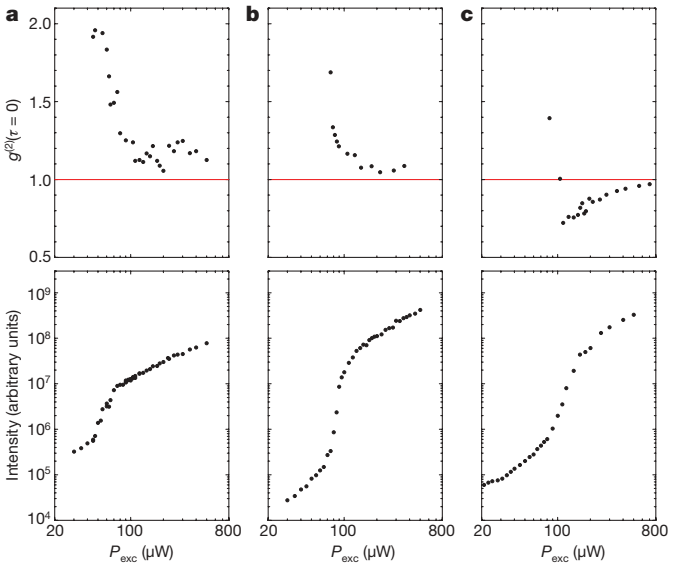


Figure 2 | Measured second-order photon correlation function at zero delay time (top) and output intensity versus input pump power, P_{exc} (bottom), for three different microcavity lasers. **a**, The II-VI cavity, which has a diameter of 1.5 μm, has a low Q factor, of 1,850, for the lowest confined photon mode. Twenty quantum dots were estimated, from the quantum-dot density and the spectral overlap of their emission with the optical mode, to be involved. **b**, **c**, For the III-V cavities of diameters 5 μm (**b**) and 8 μm (**c**), Q factors of 9,000 and 19,000 and effective quantum-dot numbers of 30 and 15 were estimated, respectively.

conventional (low-β) laser. As the power is increased, a smooth drop towards unity is seen as the lasing threshold region is passed. The low-Q III-V cavity shows a similar dependence. For the high-Q III-V cavity, however, $g^{(2)}(\tau = 0)$ unexpectedly drops to values below unity around the lasing threshold, giving clear evidence for the emission of non-classical light. This is accompanied by an even smoother input–output curve. For the III-V cavities, the correlations could not be measured well below the threshold region owing to the limited sensitivity of the streak camera for wavelengths above 900 nm.

Our experimental set-up also provides access to the second-order correlation function at finite delay times. This allows us to study the dynamical evolution of the photon correlations in semiconductor microlasers. In Fig. 3, we display the evolution of the correlations for the III-V cavities and three different excitation powers. For the low-Q cavity (Fig. 3a) and excitation below the threshold region, $g^{(2)}(\tau)$ drops smoothly from a value slightly below two towards unity within the first picoseconds. However, for delays of around 40 ps, values below unity appear. The data for the II-VI cavity are similar to those for the low-Q cavity. For the high-Q cavity (Fig. 3b) and excitation in the threshold region, pronounced long-lasting oscillations of $g^{(2)}(\tau)$ are visible. To ensure that this effect is not caused by noise or lack of statistics, we studied the power dependence of the amplitude and frequency of the oscillations and found systematic variations (not shown). The oscillations become obviously damped with increasing excitation power. Above threshold, we find that $g^{(2)}(\tau) \approx 1$ in both the low-Q case and the high-Q case.

For $g^{(2)}(0)$, a semiconductor theory has been developed recently[18] using the cluster expansion technique[19–21]. In contrast to models based on atomic systems, this theory is capable of incorporating semiconductor-specific features such as Pauli blocking, Coulomb interaction and a modified source term for spontaneous emission[22–24]. The starting point of the microscopic theory is a semiconductor Hamiltonian for the interacting carrier–photon system. The time evolution of the photon number in the laser mode, $\langle b^\dagger b\rangle$, and the electron, $f_\nu^{\mathrm{e}} = \langle c_\nu^\dagger c_\nu\rangle$, and hole, $f_\nu^{\mathrm{h}} = 1 - \langle v_\nu^\dagger v_\nu\rangle$, populations is determined by using Heisenberg's equations of motion. Here $c_\nu^\dagger$ and c_ν and $v_\nu^\dagger$ and v_ν are the creation and annihilation operators of electrons in the conduction and, respectively, valence band states ν. Cavity losses are introduced by coupling the system to an external reservoir. Scattering and dephasing are treated in the relaxation-time approximation.

Here we extend the theory to finite delay times, τ. Suppressing the dependence on t, we rewrite equation (1) as

$$G^{(2)}(\tau) = \langle\langle b^\dagger(\tau) b(\tau)\rangle\rangle \langle b^\dagger b\rangle$$

where the expectation value $\langle\langle \cdots \rangle\rangle$ is taken with respect to the modified density operator $\tilde{\rho} = b\rho b^\dagger / \langle b^\dagger b\rangle$. This way, we reduce the two-time problem to two single-time problems that can be solved successively. Our approach, which is applied to the nonlinear equations of motion describing semiconductor systems, becomes the quantum regression theorem[25] for linear systems of equations, as frequently used for two-level systems.

In Fig. 4, we show examples of the numerical calculations. They are intended to demonstrate possible results for two different sets of parameters. The data are sensitive to the microscopic description of the carrier scattering that provides a common source for carrier redistribution and dephasing. These processes depend on, among other things, the electronic states—both for the recombination processes and for where the carriers are pumped. To simplify this rather involved analysis, we assume pumping at higher quantum-dot states; for a more direct comparison with the experiments, excitation of delocalized barrier states and subsequent capture and relaxation processes should be considered. We attribute deviations (such as the stronger damping of the oscillations in Fig. 4d) to the simplifications in the quantum-dot model. Nevertheless, the chosen examples reproduce the general trends of the experiments.

The value of $g^{(2)}(0)$ calculated for the low-Q case does not reach the subthreshold value of two for thermal light, in agreement with experiment. This has been identified before as a characteristic of microcavity lasers with larger β factors and a small numbers of emitters[3]. The corresponding calculations of $g^{(2)}(\tau)$ for low excitation intensities show a non-monotonic decay to unity from the initial value, $g^{(2)}(0)$. In the higher-Q case, we find an anti-bunching effect ($g^{(2)}(0) < 1$ and oscillations in $g^{(2)}(\tau)$), as seen in the experimental data.

The observed oscillations in the photon correlations are a result of the dynamical coupling between photons and carriers and can be qualitatively understood as follows. For a microcavity operated at steady state in the spontaneous-emission regime, $g^{(2)}(\tau)$ decays from (nearly) two to one on the timescale of the coherence time. In the regime of dominating stimulated emission, $g^{(2)}(\tau)$ equals one independently of time delay. The oscillations are observed in the regime of transition from spontaneous to stimulated emission in a system that, under these conditions, contains only very few photons emitted by very few quantum dots. This is the transition regime of cavity QED lasers discussed in ref. 10. Unlike in the situation in the lasing regime, here the loss of a photon from of the cavity represents a severe perturbation of the system, which strongly influences the coupled carrier–photon dynamics.

Systems of emitters coupled to a cavity mode are known to exhibit different kinds of oscillations of the emission intensity. Relaxation oscillations can occur close to the threshold region when the laser is switched on or perturbed, and Rabi oscillations can occur in the

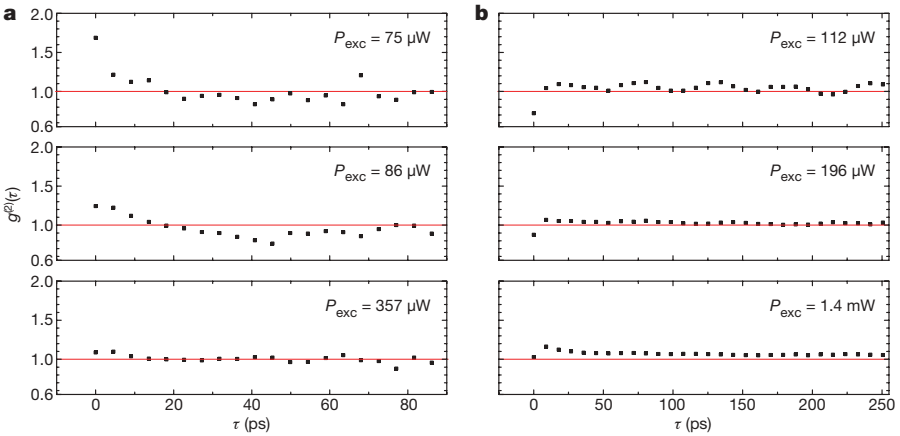


Figure 3 | Measured temporal evolution of the second-order correlation function for selected pump powers, P_{exc}. Results for the III-V cavities of Fig. 2: **a**, low-Q micropillar; **b**, high-Q micropillar. An unexpected outcome is the appearance of dynamical anti-bunching ($g^{(2)}(\tau) < 1$) for low and intermediate excitation densities at finite τ for the low-Q cavity. The intensity correlation function does not merely drop to a value of unity, but takes on smaller values with subsequent oscillations; this is particularly clear for the high-Q cavity with $P = 112$ μW.

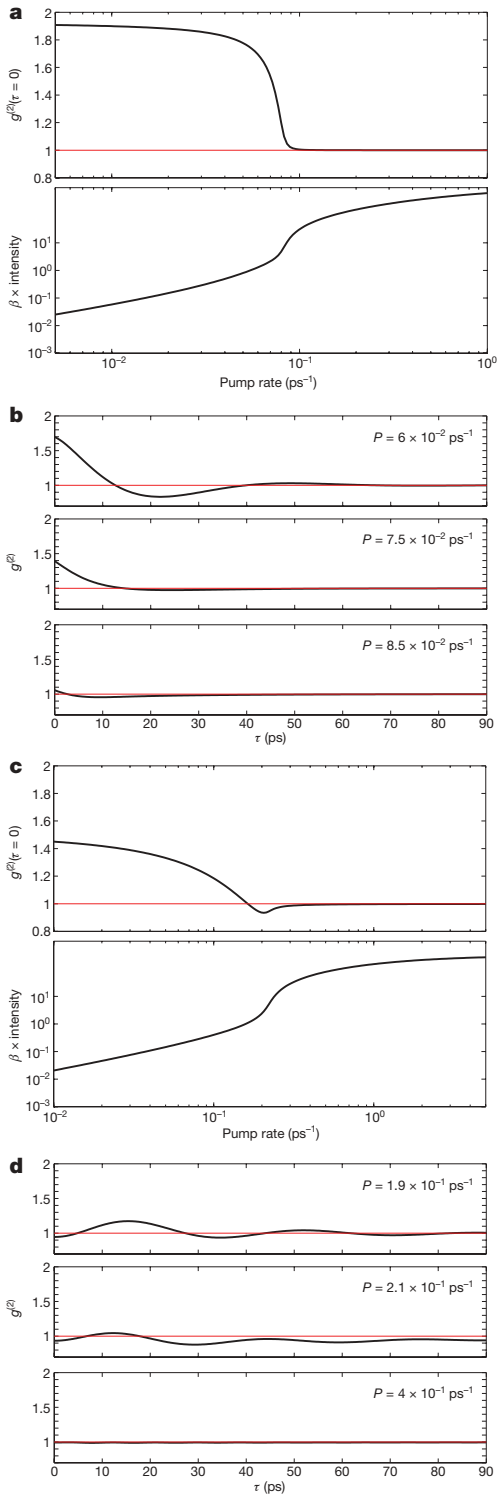


Figure 4 | Calculated zero-delay correlation functions, input–output curves and temporal dynamics of $g^{(2)}(\tau)$. a, **b**, Low-Q quantum-dot microcavity laser; **c**, **d**, high-Q quantum-dot microcavity laser. Selected pump rates, P, are used in **b** and **d**.

regime in which the dissipation is small in comparison with the light–matter coupling strength. The dynamics of the correlation function $g^{(2)}(\tau)$ can be traced back to this behaviour. Our theory predicts that both kinds of oscillations can be triggered by photon emission events. In reaction to the perturbation, the system tries to re-establish equilibrium, and, in doing so, undergoes quantum oscillations. In both cases, the origin of these oscillations is the feedback due to the cavity, which can lead to out-of-phase oscillations of photon number and lasing medium. These oscillations become damped as the pump rate increases and a regime of stimulated emission is reached in which the photon number is high enough that single photon losses no longer affect the system considerably.

In our case, the perturbation of the few-emitter system can become so prominent that, for example, the subsequent emission of a photon is suppressed. This leads to the dynamic anti-bunching both for zero delay and for times after enhanced photon pair emission during the oscillations.

These oscillations potentially also appear in intensity measurements. However, as the moment of photon emission is stochastic, any time averaging blurs the oscillations. Nevertheless, the oscillations also carry over to the correlation functions similar to $g^{(2)}(\tau)$. There they survive the averaging, as although the moment of arrival of the first photon is still stochastic, a second photon is picked whose delay, τ, relative to the first is fixed for all detected photon pairs. For increasing Q, the cavity feedback is enhanced, causing the quantum fluctuations to become more pronounced as reflected by the oscillations of $g^{(2)}(\tau)$.

In conclusion, we have presented a new measurement scheme with the necessary time resolution to fully access photon correlations on a single-photon level, yielding a record of arrival times of individual photons. From these data, correlation functions can be constructed to arbitrary order, facilitating a complete characterization of the light source. Our approach can be viewed as the most direct experimental realization of the pioneering theory in ref. 1. To give an example of these possibilities, we show results for the third-order correlation function in Fig. 5. Higher-order correlations in the light emission may provide a detailed understanding of the nature of quantum many-particle states such as those currently discussed for microcavities in the regime of strong light–matter coupling.

The validity of the reported results is not restricted to semiconductor micropillars: similar results are expected for all light sources that involve a small number of emitters, be they atoms, molecules, defects in solids or semiconductor quantum structures. We expect our findings and methods to mark the starting point for novel quantum optical studies addressing the dynamics of correlation functions of light.

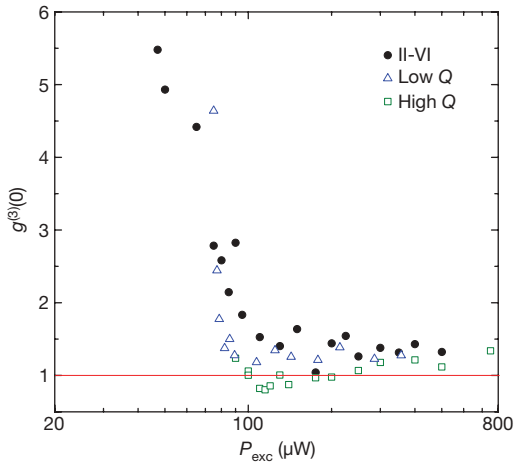


Figure 5 | Measured third-order correlation function $g^{(3)}(0)$. Characterization of the simultaneous arrival of three photons as a function of excitation power, for the three different micropillar samples studied in this paper. For the II-VI and low-Q resonators, the time resolution was increased to about 20 ps. These two samples show values close to the theoretical prediction, $g^{(3)}(0) = 6$, at lowest excitation powers. The kth-order correlation functions for a thermal source of radiation are readily obtained from the factorial moments of the Planck distribution, $\langle n(n-1)\cdots(n-k+1)\rangle = k!\langle n\rangle^k$, to give $g^{(k)}(0) = k!$, corresponding to the number of permutations of indistinguishable photons.

METHODS SUMMARY

Experimental set-up. For the intensity correlation measurements, we used a microscope objective to focus and collect the light. An interference filter with a bandwidth of 1 nm singled out the fundamental-mode emission of the microcavity. The normalized correlation function, $g^{(2)}(t,\tau)$, was determined by dividing the number of photon pairs detected using the streak camera at times t and $t+\tau$ by the product of the mean photon counting rates at the same times. Weighted averaging over all times t gave $g^{(2)}(\tau)$. To exclude noise or saturation effects, correlation measurements were also performed on a pulsed laser with several levels of attenuation, showing no deviations from the expected value, $g^{(2)}(0)=1$.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 January; accepted 28 April 2009.

1. Glauber, R. J. The quantum theory of optical coherence. *Phys. Rev.* **130,** 2529–2539 (1963).
2. Strauf, S. *et al.* Self-tuned quantum dot gain in photonic crystal lasers. *Phys. Rev. Lett.* **96,** 127404 (2006).
3. Ulrich, S. M. *et al.* Photon statistics of semiconductor microcavity lasers. *Phys. Rev. Lett.* **98,** 043906 (2007).
4. Hanbury Brown, R. & Twiss, R. Q. Correlation between photons in two coherent beams of light. *Nature* **177,** 27–29 (1956).
5. Vahala, K. J. Optical microcavities. *Nature* **424,** 839–846 (2003).
6. Choi, J.-S. *et al.* Evolution of the onset of coherence in a family of photonic crystal nanolasers. *Appl. Phys. Lett.* **91,** 031108 (2007).
7. Xie, Z. G., Götzinger, S., Fang, W., Cao, H. & Solomon, G. S. Influence of a single quantum dot state on the characteristics of a microdisk laser. *Phys. Rev. Lett.* **98,** 117401 (2007).
8. Reitzenstein, S. *et al.* Single quantum dot controlled lasing effects in high-*Q* micropillar cavities. *Opt. Express* **16,** 4848–4857 (2008).
9. Berman, P. *Cavity Quantum Electrodynamics* (Academic, 1994).
10. Rice, P. R. & Carmichael, H. J. Photon statistics of a cavity QED laser: a comment on the laser-phase-transition analogy. *Phys. Rev. A* **50,** 4318–4329 (1994).
11. Loudon, R. *The Quantum Theory of Light* 2nd edn (Clarendon, 1983).
12. Kimble, H. J., Dagenais, M. & Mandel, L. Photon antibunching in resonance fluorescence. *Phys. Rev. Lett.* **39,** 691–695 (1977).
13. Michler, P. *et al.* A quantum dot single-photon turnstile device. *Science* **290,** 2282–2285 (2000).
14. Machida, S., Yamamoto, Y. & Itaya, Y. Observation of amplitude squeezing in a constant-current-driven semiconductor laser. *Phys. Rev. Lett.* **58,** 1000–1003 (1987).
15. Hofmann, H. F. & Hess, O. Coexistence of thermal noise and squeezing in the intensity fluctuations of small laser diodes. *J. Opt. Soc. Am. B* **17,** 1926–1933 (2000).
16. Wiele, C., Haake, F., Rocke, C. & Wixforth, A. Photon trains and lasing: the periodically pumped quantum dot. *Phys. Rev. A* **58,** R2680–R2683 (1998).
17. Benson, O. & Yamamoto, Y. Master-equation model of a single-quantum-dot microsphere laser. *Phys. Rev. A* **59,** 4756–4763 (1999).
18. Gies, C., Wiersig, J., Lorke, M. & Jahnke, F. Semiconductor model for quantum dot-based microcavity lasers. *Phys. Rev. A* **75,** 013803 (2007).
19. Fricke, J. Transport equations including many-particle correlations for an arbitrary quantum system: a general formalism. *Ann. Phys.* **252,** 479–498 (1996).
20. Hoyer, W., Kira, M. & Koch, S. W. in *Nonequilibrium Physics at Short Time Scales* (ed. Morawetz, K.) 309–338 (Springer, 2004).
21. Kira, M., Jahnke, F., Hoyer, W. & Koch, S. W. Quantum theory of spontaneous emission and coherent effects in semiconductor microstructures. *Prog. Quantum Electron.* **23,** 189–279 (1999).
22. Schwab, M. *et al.* Radiative emission dynamics of quantum dots in a single cavity micropillar. *Phys. Rev. B* **74,** 045323 (2006).
23. Baer, N., Gies, C., Wiersig, J. & Jahnke, F. Luminescence of a semiconductor quantum dot system. *Eur. Phys. J. B* **50,** 411–418 (2006).
24. Berstermann, T. *et al.* Systematic study of carrier correlations in the electron-hole recombination dynamics of quantum dots. *Phys. Rev. B* **76,** 165318 (2007).
25. Carmichael, H. J. *Statistical Methods in Quantum Optics 1* (Springer, 1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank P. Gartner and I. Akimov for discussions and technical support. Funding from the Deutsche Forschungsgemeinschaft through the research group 'Quantum optics in semiconductor nanostructures' and a grant for CPU time at the Forschungszentrum Jülich (Germany) is gratefully acknowledged.

Author Contributions Experiments were performed in Dortmund by M.A., T.B. and M.B., with the participation of C. Kistner. Calculations were done in Bremen by J.W., C.G. and F.J. The III-V samples were grown in Würzburg by C. Kistner, S.R., C.S., S.H. and A.F. The II-VI sample was prepared in Bremen by C. Kruse, J.K. and D.H.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.J. (jahnke@itp.uni-bremen.de) or M.B. (manfred.bayer@tu-dortmund.de).

METHODS

Sample preparation. The cylindrical low-Q and high-Q III-V micropillar samples were grown by molecular beam epitaxy on a GaAs substrate. The distributed Bragg reflectors consisted of 20 upper and 23 lower alternating layers of AlAs (79 nm)–GaAs (67 nm) $\lambda/4$ pairs for the low-Q micropillar and 26 upper and 33 lower alternating layers of AlAs (74 nm)–GaAs (68 nm) $\lambda/4$ pairs for the high-Q micropillar. The central λ cavity contained one layer of self-assembled InGaAs quantum dots with a density of $\sim 3 \times 10^{10}\,\mathrm{cm}^{-2}$ in the low-Q case and one layer of self-assembled AlGaInAs quantum dots with a density of $\sim 6 \times 10^{9}\,\mathrm{cm}^{-2}$ in the high-Q case, from which cavities with diameters of several micrometres were fabricated by means of high-resolution electron beam lithography and plasma-induced reactive ion etching.

The II-VI sample was also grown by molecular beam epitaxy. There the distributed Bragg reflectors consisted of 15 upper and 18 bottom layers in which $ZnS_{0.06}Se_{0.94}$ (48 nm) layers were used as high-index material and a 25.5-period MgS (1.7 nm)–ZnCdSe (0.6 nm) superlattice was used as low-index material. The central λ cavity contained a single sheet of CdSe/ZnSe quantum dots with an approximate density of $\sim 5 \times 10^{10}\,\mathrm{cm}^{-2}$.

To estimate the number of quantum dots involved in the light emission, we multiplied the dot density of the as-grown sample by the effective photon mode area in the pillar plane. Then we scaled this number by the spectral overlap between the inhomogeneous dot emission and the width of the mode in the lasing regime. Below threshold, the spectral width of the mode is larger, but simultaneously fewer quantum dots are excited by the laser. Therefore, for simplicity we used the same number of quantum dots over the whole excitation power range.

Experimental set-up. The sample was mounted in a helium-flow microscopy cryostat, in which temperatures as low as 6 K could be achieved. A Ti:sapphire laser operating at $\lambda = 780$ nm with pulse durations of $\sim$100 fs and a repetition frequency of 76 MHz was used for excitation, corresponding to 13.2-ns pulse separation. A microscope objective with a focal length of 2 cm and a numerical aperture of 0.26 was used to focus and collect the light. After collection, the emission was directed into the streak camera. An interference filter with a bandwidth of 1 nm was used to single out the fundamental-mode emission of the microcavity. To ensure that the measured correlations did not arise as a result of noise in the charge-coupled device or saturation effects, we performed additional correlation measurements on an attenuated pulsed laser. We varied the pulsed laser intensity over the whole range of intensities used in this paper, but no deviations from the expected result ($g^{(2)}(0) = 1$) were found.

Obtaining $g^{(2)}$ from experimental data. The normalized correlation function, $g^{(2)}(t, \tau)$, is the number of photon pairs detected at times t and $t + \tau$ divided by the number of photon pairs that would be detected if the photons had the same temporal intensity profile but were emitted statistically independently of each other. The number of detected photon pairs is thus directly available in the recorded single screens of the streak camera. This number is equivalent to the non-normalized correlation function, $G^{(2)}(t, \tau)$, which is usually measured using a Hanbury Brown/Twiss set-up. In the case of statistically independent photons, the information about the number of expected photon pairs is completely determined by the mean intensities at times t and $t + \tau$. Therefore, the normalization of $g^{(2)}(t, t + \tau)$ can be calculated from the intensity profile integrated over all screens. Summing all photon pairs at various times, t, and for a fixed delay, τ, and dividing this number by the sum of the expected statistically independent photon pairs for variable t values and fixed τ, gives the time-averaged function $g^{(2)}(\tau)$. In the excitation density regime of the transition from spontaneous to stimulated emission, and above, the photon pairs at late t (when the emission is purely spontaneous) have a negligible influence on the photon statistics.

Quantum-dot parameters. In the numerical calculations, we assumed that 50 quantum dots in the low-Q micropillar and eight quantum dots in the high-Q micropillar were resonant with the optical mode. We chose a spontaneous-emission factor of $\beta = 0.1$, a total spontaneous-emission time (enhanced as a result of the Purcell effect) of $\tau_{sp} = 1.7$ ps (low Q) or 0.75 ps (high Q), and a quality factor of $Q = 40{,}000$ (low) or 80,000 (high). We note that the experimental Q factors are lower bounds, as they are determined for weak excitation below the transparency point. Considering resonant pumping of carriers in the quantum-dot p shell, we used a relaxation time from p shell to s shell of 0.5 ps for electrons and 0.25 ps for holes. Further details of the quantum-dot model are given in ref. 23.

Details of calculations. Coupled equations of motion for various expectation values and correlation functions can be obtained[18] using Heisenberg's equations of motion for the carrier and photon operators. To truncate the arising hierarchy of equations, the cluster expansion method[19,20] was used. In this approach, operator averages are classified into singlets, doublets, triplets, quadruplets, etc., according to the number of particles they involve. Mixed expectation values with carrier and photon operators are classified accordingly, keeping in mind that a photon operator is linked to two carrier operators[21]. Truncation on the doublet level has been used to describe photoluminescence from quantum dots[23]. To compute the photon statistics in terms of $g^{(2)}(0)$, truncation at the quadruplet level is required[18].

First we computed $\langle b^\dagger(t)b(t)\rangle$, $\langle b^\dagger(t)b^\dagger(t)b(t)b(t)\rangle$, $\langle b^\dagger(t)b(t)c^\dagger(t)c(t)\rangle$, … with $t \to \infty$ according to the single-time equations of motion on quadruplet level as discussed in ref. 18. In a second step, the τ dynamics of

$$G^{(2)}(\tau) = \langle\langle b^\dagger(\tau)b(\tau)\rangle\rangle\langle b^\dagger b\rangle$$

was evaluated using the single-time equations of motion for $\langle\langle b^\dagger(\tau)b(\tau)\rangle\rangle$, $\langle\langle c_v^\dagger(\tau)c_v(\tau)\rangle\rangle$, … with the initial conditions

$$\langle\langle b^\dagger(\tau)b(\tau)\rangle\rangle|_{\tau=0} = \langle b^\dagger b^\dagger bb\rangle/\langle b^\dagger b\rangle$$

$$\langle\langle c_v^\dagger(\tau)c_v(\tau)\rangle\rangle|_{\tau=0} = \langle b^\dagger bc_v^\dagger c_v\rangle/\langle b^\dagger b\rangle$$

and so on. The underlying assumption is that the truncation introduced by the cluster expansion works equally well for operator averages taken with respect to the density operators ρ and $\tilde{\rho}$.

LETTERS

DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes

Xiaomin Tu[1], Suresh Manohar[2], Anand Jagota[2,3] & Ming Zheng[1]

Single-walled carbon nanotubes (SWNTs) are a family of molecules that have the same cylindrical shape but different chiralities[1]. Many fundamental studies and technological applications[2] of SWNTs require a population of tubes with identical chirality that current syntheses cannot provide. The SWNT sorting problem—that is, separation of a synthetic mixture of tubes into individual single-chirality components—has attracted considerable attention in recent years. Intense efforts so far have focused largely on, and resulted in solutions for, a weaker version of the sorting problem: metal/semiconductor separation[3,4]. A systematic and general method to purify each and every single-chirality species of the same electronic type from the synthetic mixture of SWNTs is highly desirable, but the task has proven to be insurmountable to date. Here we report such a method, which allows purification of all 12 major single-chirality semiconducting species from a synthetic mixture, with sufficient yield for both fundamental studies and application development. We have designed an effective search of a DNA library of ~10^{60} in size, and have identified more than 20 short DNA sequences, each of which recognizes and enables chromatographic purification of a particular nanotube species from the synthetic mixture. Recognition sequences exhibit a periodic purine–pyrimidines pattern, which can undergo hydrogen-bonding to form a two-dimensional sheet, and fold selectively on nanotubes into a well-ordered three-dimensional barrel. We propose that the ordered two-dimensional sheet and three-dimensional barrel provide the structural basis for the observed DNA recognition of SWNTs.

We have previously shown that a particular DNA sequence, poly(GT), forms ordered structures on SWNTs[5], allowing low-resolution metal/semiconductor and diameter sorting by ion exchange chromatography (IEX). Under favourable conditions, a few small diameter single-chirality species can be purified[6]. These results have prompted many experimental[5,7,8] and theoretical investigations[9–15] on DNA–SWNT interactions. Emerging from these studies is a notion that DNA–SWNT interaction and the resulting hybrid structure are dependent on both DNA sequence and SWNT structure. This points to the possibility of selecting DNA sequences for specific SWNT structure recognition. The number of unique sequences available from current solid phase oligonucleotide synthesis is at least 4^{100} or 10^{60}. The sheer size of this molecular library suggests that it is highly probable that specific SWNT recognition sequences exist, but the chance of finding these sequences through random sampling of the library could be exceedingly low.

To efficiently and systematically explore the DNA library, we first devised a sequence pattern expansion scheme to coarsely survey the library in a confined and progressive manner. In the first-order expansion, all four possible mononucleotide repeats such as TTT… are examined; in the second-order expansion, all 16 possible di-nucleotide repeats such as ATATAT… are examined. Likewise, in the *n*th order expansion, all *n*-nucleotide repeats are examined. The examination process involves two steps (see Supplementary Information for experimental details). Each DNA sequence is first used to solubilize, through ultrasonic dispersion, a synthetic mixture of tubes made by the HiPco process[7]. The dispersion is then fractionated by IEX[3,5]. As successful fractionation requires well-defined DNA–SWNT structures, the chromatography step essentially serves as an assay to select sequences that can form well-ordered structures on SWNTs.

In this work, we carried out the expansion to the fourth order. To further reduce the number of test sequences to a manageable level, we initially fixed sequence length to 28- or 30-mers, and ignored the effect of sequence ends such that patterns like $(ATT)_{10}$, $(TTA)_{10}$ and $(TAT)_{10}$ are considered to be identical. Although most sequences tested in the survey do not yield any substantial fractionation of the starting material, a set of sequences with a pattern of alternating purine–pyrimidine(s) does provide varying degrees of separation. These include previously identified $(GT)_{15}$, $(TCG)_{10}$ and $(ATTT)_7$, among others. For instance, IEX separation of $(TCG)_{10}$ dispersed HiPco gives larger diameter semiconductor tubes in the early fractions, whereas metallic and small diameter semiconductor tubes are eluted at a later stage. Similarly, the four-nucleotide repeat $(ATTT)_7$ allows clean enrichment of larger diameter semiconducting tubes in the early fractions. These elution profiles are completely opposite to that of the poly(GT) dispersed HiPco, which gives metallic and small diameter semiconducting tubes first, followed by large diameter semiconducting tubes[5]. The poly(GT)–SWNT elution profile was explained by a model that invokes only electrostatic interaction between DNA–SWNT and IEX resin[5,16]. To account for the elution profiles observed with newly identified sequences, it is very likely that interactions other than electrostatic ones need to be considered in determining the equilibrium binding between IEX resin and SWNT dispersion.

To refine the coarse survey results, we then studied the effect of length variation on the sequence patterns identified by the expansion scheme. Early studies on the poly(GT) sequence showed little length effect on IEX separation for chains longer than 20-mers[5]. A surprising finding in this study is that when sequence length is reduced to ~10-mers, such that the backbone contour length is comparable to the van der Waals circumference of a typical SWNT 1 nm in diameter, the IEX separation becomes extremely sensitive to the length of the DNA strand. For example, the 12-mer $(GT)_6$ allows early elution and purification of (8,6) (for SWNT atomic structure notation, see ref. 1) with little metallic contaminants; a sequence just one base shorter, $(GT)_5G$, enriches predominantly one metallic tube in the early fractions. Likewise, the 8-mer $(CCG)_2CC$ yields early elution of (8,7) tubes, whereas the 12-mer $(CCG)_4$ causes (9,4) tubes to elute first. The effect of sequence ends also becomes prominent in short

[1]DuPont Central Research and Development, Wilmington, Delaware 19880, USA. [2]Department of Chemical Engineering and [3]Bioengineering Program, Lehigh University, Bethlehem, Pennsylvania 18015, USA.

Table 1 | DNA sequence versus SWNT chirality

Chirality (*n*,*m*)	Sequences
(9,1)	$(TCC)_{10}$, $(TGA)_{10}$, $(CCA)_{10}$
(8,3)	$(TTA)_4TT$, $(TTA)_3TTGTT$, $(TTA)_5TT$
(6,5)	$(TAT)_4$, $(CGT)_3C$
(7,5)	$(ATT)_4$, $(ATT)_4AT$
(10,2)	$(TATT)_2TAT$
(8,4)	$(ATTT)_3$
(9,4)	$(GTC)_2GT$, $(CCG)_4$
(7,6)	$(GTT)_3G$, $(TGT)_4T$
(8,6)	$(GT)_6$, $(TATT)_3T$, $(TCG)_{10}$, $(GTC)_3$, $(TCG)_2TC$, $(TCG)_4TC$, $(GTC)_2$
(9,5)	$(TGTT)_2TGT$
(10,5)	$(TTTA)_3T$
(8,7)	$(CCG)_2CC$

DNA sequences enabling chromatographic purification of single chirality semiconducting SWNTs.

sequences: $(TAT)_4$ enables early elution of (6,5), but its compositional isomer $(ATT)_4$ enables that of (7,5). Prompted by these findings, we examined most of the 8- to 16-mer sequences that conform to the alternating purine-pyrimidine(s) pattern.

From a total of ~350 sequences tested, we identified more than 20 sequences (Table 1), each of which allows purification of a particular (*n*,*m*) species. The exquisite selectivity of DNA is made clear by its ability to purify low abundance species such as (9,1) and (8,3), and by the need for different sequences to purify two tubes, (9,1) and (6,5), that have identical diameter. For reasons not entirely clear to us, semiconducting tubes are much more likely to be purified. We speculate that metallic tubes with different (*n*,*m*) values share similar electronic character, and they are less sensitive to perturbation by molecular binding than semiconducting tubes. As a result, the interactions of metallic tubes with DNA show less dependence on tube structures. Supplementary Fig. 1 shows a typical IEX elution profile, in which (8,4) tubes are eluted at the earliest stage among all tube species. Figure 1 compiles optical absorption spectra of all 12 purified semiconducting (*n*,*m*) species, along with that of the starting material. Each spectrum of the purified species exhibits a series of sharp absorption peaks corresponding to the E_{11}, E_{22} and E_{33} transitions. For (8,3), (9,4), (7,6), (8,6) and (8,7) tubes, even E_{44} can be clearly identified. Supplementary Table 1 lists the E_{ii} peak positions in each spectrum, and the corresponding (*n*,*m*) assignment according to values derived from fluorescence measurements[17,18]. The 12 purified species represent all the major semiconducting components of the starting material. This conclusion is based on fluorescence mapping (Supplementary Fig. 2), and fitting of the absorption spectrum (Supplementary Fig. 3.2) of the starting material.

The absorption spectrum for each (*n*,*m*) fraction in Fig. 1 shows a different level of contamination by metallic and other semiconducting species. We have estimated this level (Supplementary Information) according to a method developed by others[19]. The results are listed in

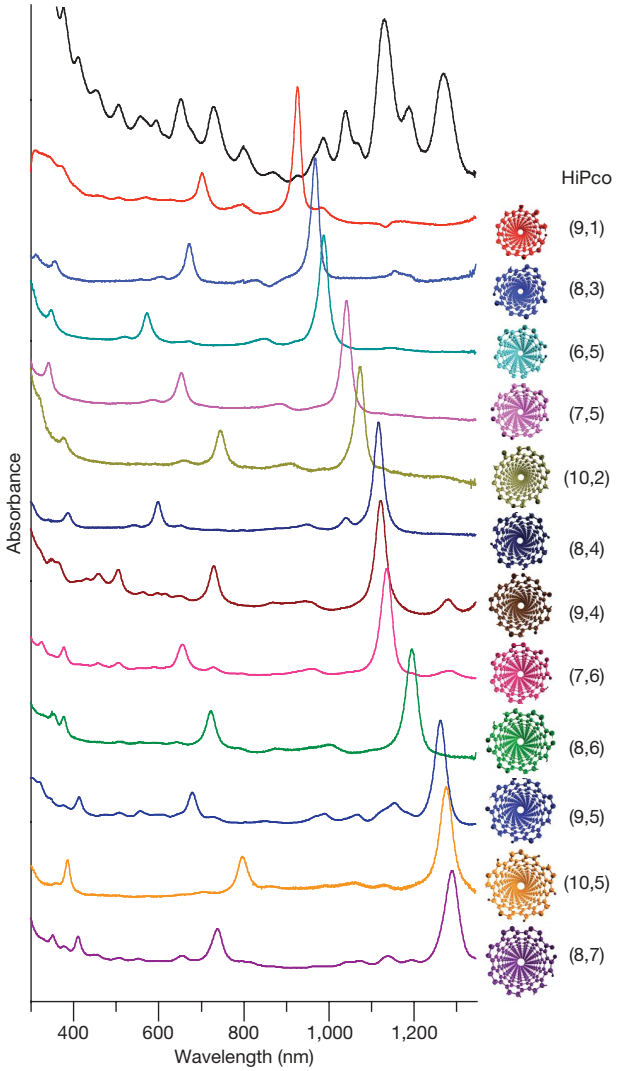


Figure 1 | Optical absorption spectra and atomic structures. Ultraviolet–visible–near-infrared absorption spectra of 12 purified semiconducting SWNTs (ranked according to the measured E_{11} absorption wavelength) and the starting HiPco mixture. The structure of each purified SWNT species (viewed along the tube axis) and its (*n*,*m*) notation are given at the right side of the corresponding spectrum. Each purified SWNT was obtained according to procedures described in Supplementary Information. The elution buffer was replaced by de-ionized water before spectral measurement. The baseline level of each spectrum was offset to facilitate visual comparison.

Table 2 | Experimental conditions and quantification

Chirality (*n*,*m*)	Sequence	Dispersion solution*	Incubation period†	Yield‡ (μg per 100 μg)	Purity§ (%)
(9,1)	$(TCC)_{10}$	0.1 M NaCl	None	0.1	80
(8,3)	$(TTA)_3TTGTT$	0.1 M NaCl	1 day	0.1	70
(6,5)	$(TAT)_4$	0.1 M NaCl	None	0.5	90
(7,5)	$(ATT)_4AT$	0.1 M NaCl	None	0.2	90
(10,2)	$(TATT)_2TAT$	0.1 M NaCl	None	0.1	90
(8,4)	$(ATTT)_3$	0.1 M NaCl	None	0.3	90
(9,4)	$(GTC)_2GT$	0.1 M sodium acetate pH 4.5	2 days	0.5	60
(7,6)	$(GTT)_3G$	0.1 M NaCl	None	0.4	90
(8,6)	$(GT)_6$	0.1 M NaCl	1 day	0.8	90
(9,5)	$(TGTT)_2TGT$	0.1 M NaCl, 10% glycerol	None	0.3	70
(10,5)	$(TTTA)_3T$	0.1 M sodium acetate pH 4.5	2 days	0.5	90
(8,7)	$(CCG)_2CC$	0.1 M NaCl, 10% glycerol	None	0.4	80

The experimental conditions listed here are used to obtain the results shown in Fig. 1.
* Solution in which DNA and raw HiPco material are mixed and sonicated.
† Time period between the end of sonication dispersion and the beginning of the IEX separation.
‡ The mass of (*n*,*m*) tubes obtained from 100 μg SWNTs loaded on to the IEX column in each run.
§ Estimated error ±10%, due to extinction coefficient difference among different semiconducting tubes, the presence of baseline and phonon side-bands near E_{11} peaks[28].

Table 2. We note the approximate nature of these numbers, as phonon side-bands and a difference in extinction coefficients among semiconducting tubes are neglected. To obtain an independent check on the purity estimation, we have employed direct electronic measurement on the (10,5) tubes. Field effect transistors with on/off ratios as high as 10^6 have been achieved with the (10,5) enriched fraction, and the device success rate suggests that 99% of the tubes in the fraction are semiconducting[20]. This result is consistent with the purity estimate for the (10,5) fraction based on spectral analysis.

We have also estimated purification yield for each (*n*,*m*) tube by absorption spectral analysis (Supplementary Information). The results are given in Table 2. In addition to the choice of DNA sequences, we have also identified some other factors that affect the yield. For example, our initial purification of (10,5) used 0.1 M NaCl as the default SWNT dispersion solution. Under this condition, ~0.1 μg of the (10,5) species were purified. Subsequently, we found that the chemical composition and pH of the dispersion solution, as well as the incubation time given to the dispersed tubes before loading them onto the IEX column, strongly affect the yield. After optimization, the (10,5) yield was increased by fivefold to ~0.5 μg, corresponding to about 10% of the (10,5) tubes contained in the starting material. Table 2 lists the purification conditions for some (*n*,*m*) tubes we have optimized so far. Other factors limiting the yield include irreversible SWNT adsorption on to the IEX resin, and finite resolution of the elution profile causing mixing-in of other species (Supplementary Fig. 1). The yield of a particular (*n*,*m*) species can also be dramatically improved if it is present at high level in the starting material. For instance, the yield of (6,5) species is increased by as much as 50 times when the CoMoCAT material[21] is used. Although there is no conceptual difficulty in scaling up our purification method, and we have already explored some associated engineering and cost problems[3], material cost issues make it unlikely to be realized at present. A more economically attractive option in the future is to use our process to produce high purity SWNT 'seeds' as templates for synthetic growth of the same chirality tubes. The latter 'cloning' concept has been demonstrated recently[22].

What is the structural basis for the observed DNA sequence specificity in SWNT purification? Although the answer is not completely clear to us at the moment, we do wish to point out some prominent features of the identified recognition sequences and propose a recognition mechanism. First, most of the recognition sequences in Table 1 are derived from simple pyrimidine repeats such as …TTTT…, …CCCC… and …TCTC…, with periodic purine (G or A) insertions. Second, we find that this pattern of sequences can form a stable, well ordered two-dimensional (2D) sheet through hydrogen bonding interactions between adjacent strands (Fig. 2a), resembling the well-known protein β-sheet motif. Third, we also find that the 2D sheet can be rolled up onto a particular SWNT to form a stable barrel (Fig. 2b and c, and Supplementary Information), resulting in a structure analogous to the well-known β-barrel proteins derived from protein β-sheets[23]. We propose that an ordered DNA–SWNT structure would minimize its van der Waals and hydrophobic interactions with the IEX resin, allowing it to be eluted early and purified (Supplementary Information and Supplementary Fig. 8). We also hypothesize that for each of the recognition sequences, an ordered DNA barrel structure forms only on one particular (*n*,*m*) tube, resulting in their purification by IEX.

The foregoing discussion emphasizes structural order rather than binding affinity of a DNA–SWNT hybrid. Some alternative SWNT sorting approaches based on binding affinity difference provide contrasting examples to illustrate this point. Certain aromatic monomers and polymers have been shown to be effective in selectively solubilizing semiconducting tubes of near armchair chiralities[24–27]. This method is conceptually different from ours, as DNA recognition sequences that we have identified hardly show any selection in SWNT solubilization. Rather, the selection comes from the electrostatic and electrodynamic interactions between DNA–SWNT hybrids and ion exchange resin. Whether or not the ordered DNA structures we propose here bear any biological relevance is a question we will address in future studies.

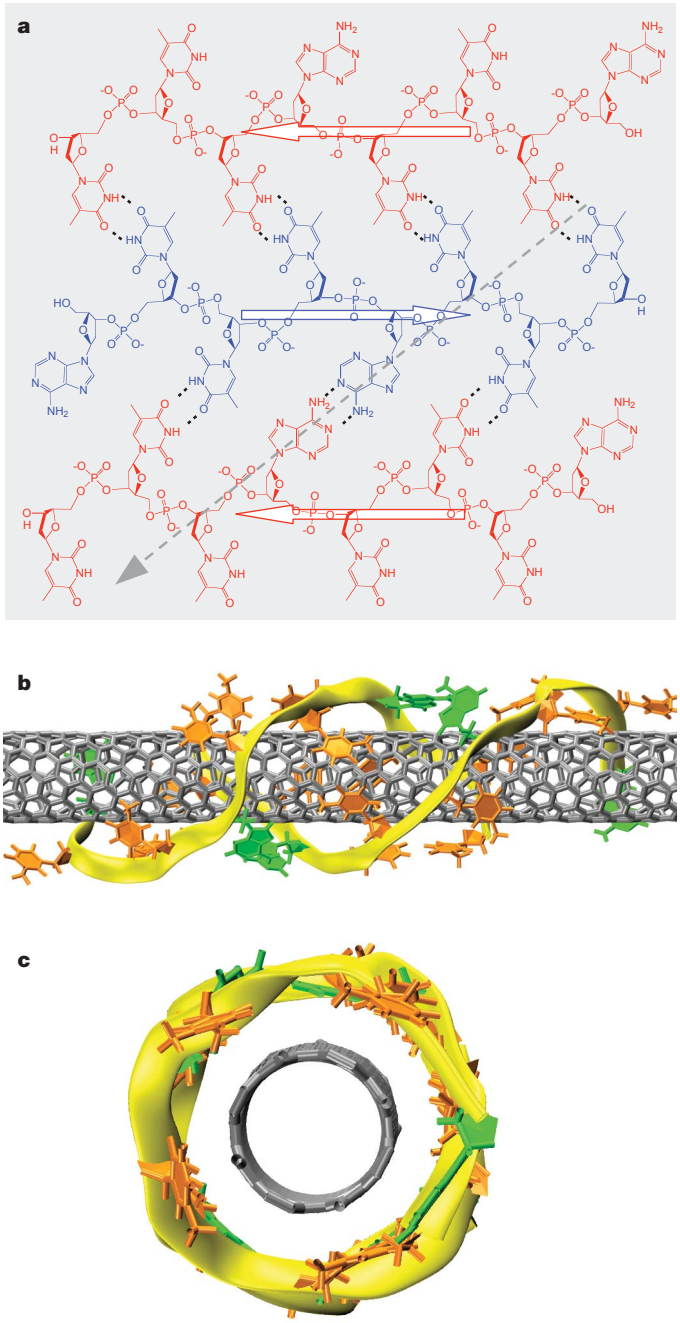


Figure 2 | DNA structures. **a**, A 2D DNA sheet structure formed by three anti-parallel ATTTATTT strands. The dotted lines between bases indicate hydrogen bonds. The open arrow in each strand denotes 5′ to 3′ direction. The dashed grey arrow (top right to bottom left) represents the roll-up vector along which the DNA barrel in **b** is formed. **b**, A DNA barrel on a (8,4) nanotube formed by rolling up a 2D DNA sheet composed of two hydrogen-bonded anti-parallel ATTTATTTATTT strands. **c**, The structure in **b** viewed along the tube axis. Colour coding: orange, thymine; green, adenine; yellow ribbons, backbones.

METHODS SUMMARY

Dispersion of SWNTs by DNA and subsequent separation by IEX follows procedures described before[5,7]. SWNT fluorescence mapping and absorption spectral analysis of the HiPco starting material, and molecular modelling of DNA–SWNT structures are performed according to methods described in Supplementary Information.

Received 5 February; accepted 1 May 2009.

1. Saito, R., Dresselhaus, G. & Dresselhaus, M. S. *Physical Properties of Carbon Nanotubes* (Imperial College Press, 1999).
2. Baughman, R. H., Zakhidov, A. A. & de Heer, W. A. Carbon nanotubes-the route toward applications. *Science* **297,** 787–792 (2002).
3. Tu, X. & Zheng, M. A DNA-based approach to the carbon nanotube sorting problem. *Nano Res.* **1,** 185–194 (2008).
4. Hersam, M. C. Progress towards monodisperse single-walled carbon nanotubes. *Nature Nanotechnol.* **3,** 387–394 (2008).
5. Zheng, M. *et al.* Structure-based carbon nanotube sorting by sequence-dependent DNA assembly. *Science* **302,** 1545–1548 (2003).
6. Zheng, M. & Semke, E. D. Enrichment of single chirality carbon nanotubes. *J. Am. Chem. Soc.* **129,** 6084–6085 (2007).
7. Zheng, M. *et al.* DNA-assisted dispersion and separation of carbon nanotubes. *Nature Mater.* **2,** 338–342 (2003).
8. Manohar, S. *et al.* Peeling single-stranded DNA from graphite surface to determine oligonucleotide binding energy by force spectroscopy. *Nano Lett.* **8,** 4365–4372 (2008).
9. Meng, S., Maregakis, P., Papaloukas, C. & Kaxiras, E. DNA nucleoside interaction and identification with carbon nanotubes. *Nano Lett.* **7,** 45–50 (2007).
10. Frischknecht, A. L. & Martin, M. G. Simulation of the adsorption of nucleotide monophosphates on carbon nanotubes in aqueous solution. *J. Phys. Chem. C* **112,** 6271–6278 (2008).
11. Gowtham, S., Scheicher, R. H., Pandey, R., Karna, S. P. & Ahuja, R. First-principles study of physisorption of nucleic acid bases on small-diameter carbon nanotubes. *Nanotechnology* **19,** 125701 (2008).
12. Johnson, R. R., Johnson, A. T. C. & Klein, M. L. Probing the structure of DNA-carbon nanotube hybrids with molecular dynamics. *Nano Lett.* **8,** 69–75 (2008).
13. Manohar, S., Tang, T. & Jagota, A. Structure of homopolymer DNA-CNT hybrids. *J. Phys. Chem. C* **111,** 17835–17845 (2007).
14. Martin, W., Zhu, W. & Krilov, G. Simulation study of noncovalent hybridization of carbon nanotubes by single-stranded DNA in water. *J. Phys. Chem. B* **112,** 16076–16089 (2008).
15. Johnson, R. R., Kohlmeyer, A., Johnson, A. T. C. & Klein, M. L. Free energy landscape of a DNA-carbon nanotube hybrid using replica exchange molecular dynamics. *Nano Lett.* **9,** 537–541 (2009).
16. Lustig, S. R., Jagota, A., Khripin, C. & Zheng, M. Theory of structure-based carbon nanotube separations by ion-exchange chromatography of DNA/CNT hybrids. *J. Phys. Chem. B* **109,** 2559–2566 (2005).
17. Bachilo, S. M. *et al.* Structure-assigned optical spectra of single-walled carbon nanotubes. *Science* **298,** 2361–2366 (2002).
18. Haroz, E. H., Bachilo, S. M., Weisman, R. B. & Doorn, S. K. Curvature effects on the E33 and E44 exciton transitions in semiconducting single-walled carbon nanotubes. *Phys. Rev. B* **77,** 125405 (2008).
19. Kim, W.-J. *et al.* Connecting single molecule electrical measurements to ensemble spectroscopic properties for quantification of single-walled carbon nanotube separation. *J. Am. Chem. Soc.* **131,** 3128–3129 (2009).
20. Zhang, L. *et al.* Optical characterizations and electronic devices of nearly pure (10,5) single-walled carbon nanotubes. *J. Am. Chem. Soc.* **131,** 2454–2455 (2009).
21. Bachilo, S. M. *et al.* Narrow (n,m)-distribution of single-walled carbon nanotubes grown using a solid supported catalyst. *J. Am. Chem. Soc.* **125,** 11186–11187 (2003).
22. Yao, Y., Feng, C., Zhang, J. & Liu, Z. "Cloning" of single-walled carbon nanotubes via open-end growth mechanism. *Nano Lett.* **9,** 1673–1677 (2009).
23. Stryer, L. *Biochemistry* (Freeman and Co., 1995).
24. Nish, A., Hwang, J., Doig, J. & Nicholas, R. J. Highly selective dispersion of single-walled carbon nanotubes using aromatic polymers. *Nature Nanotechnology* **2,** 640–646 (2007).
25. Chen, F., Wang, B., Chen, Y. & Li, L.-J. Toward the extraction of single species of single-walled carbon nanotubes using fluorene-based polymers. *Nano Lett.* **7,** 3013–3017 (2007).
26. Ju, S., Doll, J., Sharma, I. & Papadimitrakopoulos, F. Selection of carbon nanotubes with specific chiralities using helical assemblies of flavin mononucleotide. *Nature Nanotechnol.* **3,** 356–362 (2008).
27. Marquis, R. *et al.* Supramolecular discrimination of carbon nanotubes according to their helicity. *Nano Lett.* **8,** 1830–1835 (2008).
28. Zheng, M. & Diner, B. A. Solution redox chemistry of carbon nanotubes. *J. Am. Chem. Soc.* **126,** 15490–15494 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported in part by the US National Science Foundation (grant CMS-060950). We thank T. Devine for technical assistance.

Author Contributions X.T. conducted all the separation experiments and participated in their design with M.Z.; S.M. and A.J. conducted DNA–SWNT structure analysis under direction from A.J.; all authors contributed to the manuscript writing; and M.Z. guided all aspects of the work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.Z. (ming.zheng@usa.dupont.com).

Vol 460|9 July 2009|doi:10.1038/nature08141

LETTERS

Late Cretaceous seasonal ocean variability from the Arctic

Andrew Davies[1]†, Alan E. S. Kemp[1] & Jennifer Pike[2]

The modern Arctic Ocean is regarded as a barometer of global change and amplifier of global warming[1] and therefore records of past Arctic change are critical for palaeoclimate reconstruction. Little is known of the state of the Arctic Ocean in the greenhouse period of the Late Cretaceous epoch (65–99 million years ago), yet records from such times may yield important clues to Arctic Ocean behaviour in near-future warmer climates. Here we present a seasonally resolved Cretaceous sedimentary record from the Alpha ridge of the Arctic Ocean. This palaeo-sediment trap provides new insight into the workings of the Cretaceous marine biological carbon pump. Seasonal primary production was dominated by diatom algae but was not related to upwelling as was previously hypothesized[2]. Rather, production occurred within a stratified water column, involving specially adapted species in blooms resembling those of the modern North Pacific subtropical gyre[3], or those indicated for the Mediterranean sapropels[4]. With increased CO_2 levels and warming currently driving increased stratification in the global ocean[5], this style of production that is adapted to stratification may become more widespread. Our evidence for seasonal diatom production and flux testify to an ice-free summer, but thin accumulations of terrigenous sediment within the diatom ooze are consistent with the presence of intermittent sea ice in the winter, supporting a wide body of evidence for low temperatures in the Late Cretaceous Arctic Ocean[6–8], rather than recent suggestions of a 15 °C mean annual temperature at this time[9].

The Arctic is a critical yet under-sampled region for palaeoclimate studies. The recent Integrated Ocean Drilling Program Arctic Coring Expedition has provided Arctic records back to the late Palaeogene (56 Myr ago)[10] but earlier shallow coring of older, Cretaceous sediments has hitherto offered only tantalising indications of the Arctic palaeoenvironment[2,9,11,12]. Such sediments provide an opportunity to investigate Arctic climate variability in past greenhouse states that may be analogues for the future. Specifically, little is known of Arctic seasonal-scale climate variability in periods without permanent sea ice cover.

Cretaceous laminated sediments also provide a palaeo-sediment trap record with which the past workings of the marine biological carbon pump may be elucidated. In the modern ocean, diatom algae are responsible for up to 40% of oceanic primary production, and because they dominate export in many marine environments, diatoms are the key agents in the marine biological carbon pump, central to biogeochemical cycling[13,14]. In contrast, the role of diatoms in the Cretaceous oceans is poorly understood, in part owing to lack of preservation—opal A is unstable and diatoms are easily destroyed during sediment burial and silica diagenesis. However, the first, albeit rare, pelagic diatomites occur in the late Cretaceous, coinciding with a radiation of planktonic diatoms[15], and in late Cretaceous sediments without surviving biosilica, there is increasing biomarker evidence of diatom contribution to carbon cycling, including the production that generated the black shale horizons of this period[16].

With remarkable serendipity, successive US and Canadian expeditions that occupied floating ice islands above the Alpha ridge of the Arctic Ocean recovered cores containing shallow buried upper Cretaceous diatom ooze with superbly preserved diatoms (cores Fl-437 and CESAR-6 respectively; taken some 160 km apart) (Figs 1, 2)[2,11]. A third core (Fl-533) contains organic carbon-rich mud with no biosilica but with biomarker evidence for the presence of rhizosolenid diatoms[9].

Detailed diatom micropalaeontology of the material we analysed from the CESAR-6 core shows no evidence for evolutionary biostratigraphic change, and the diatom floras show closest correlation to the latest Campanian *Stephanopyxis simonseni* zone of the recently

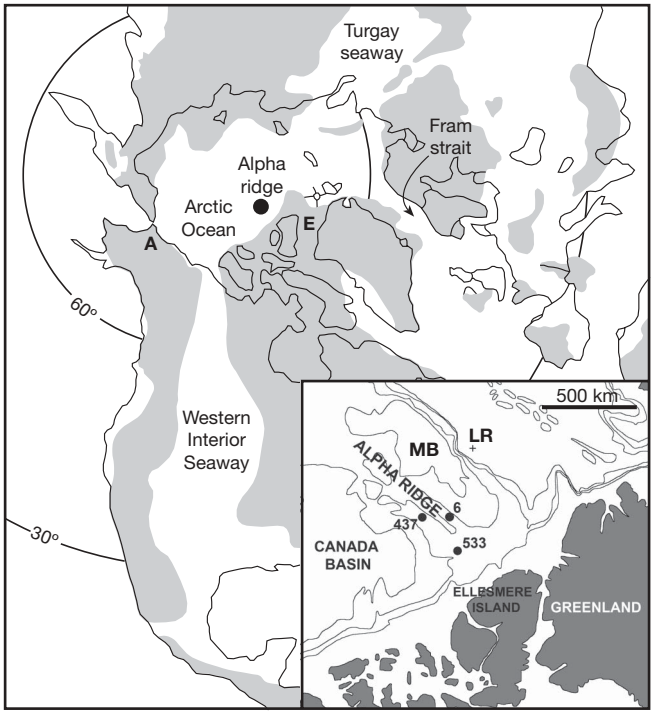


Figure 1 | Location of the Alpha ridge cores. The main panel shows Late Cretaceous palaeogeography of the Arctic region, with the location of the Alpha ridge indicated by a black circle. The other palaeoclimate sites referred to in the text are also shown: Ellesmere Island[6] (E) and Alaska location (A)[7]. The inset shows the modern geography and location of the Alpha ridge cores: 6, CESAR-6; 437, Fl-437; 533, Fl-533. LR, Lomonosov ridge; MB, Makarov basin. Adapted from the Turonian reconstruction of ref. 9.

[1]National Oceanography Centre Southampton, School of Ocean and Earth Science, University of Southampton, Southampton, SO14 3ZH, UK. [2]School of Earth and Ocean Sciences, Cardiff University, Main Building, Park Place, Cardiff, CF10 3YE, UK. †Present address: Neftex Petroleum Consultants Ltd, 97 Milton Park, Abingdon, OX14 4RY, UK.

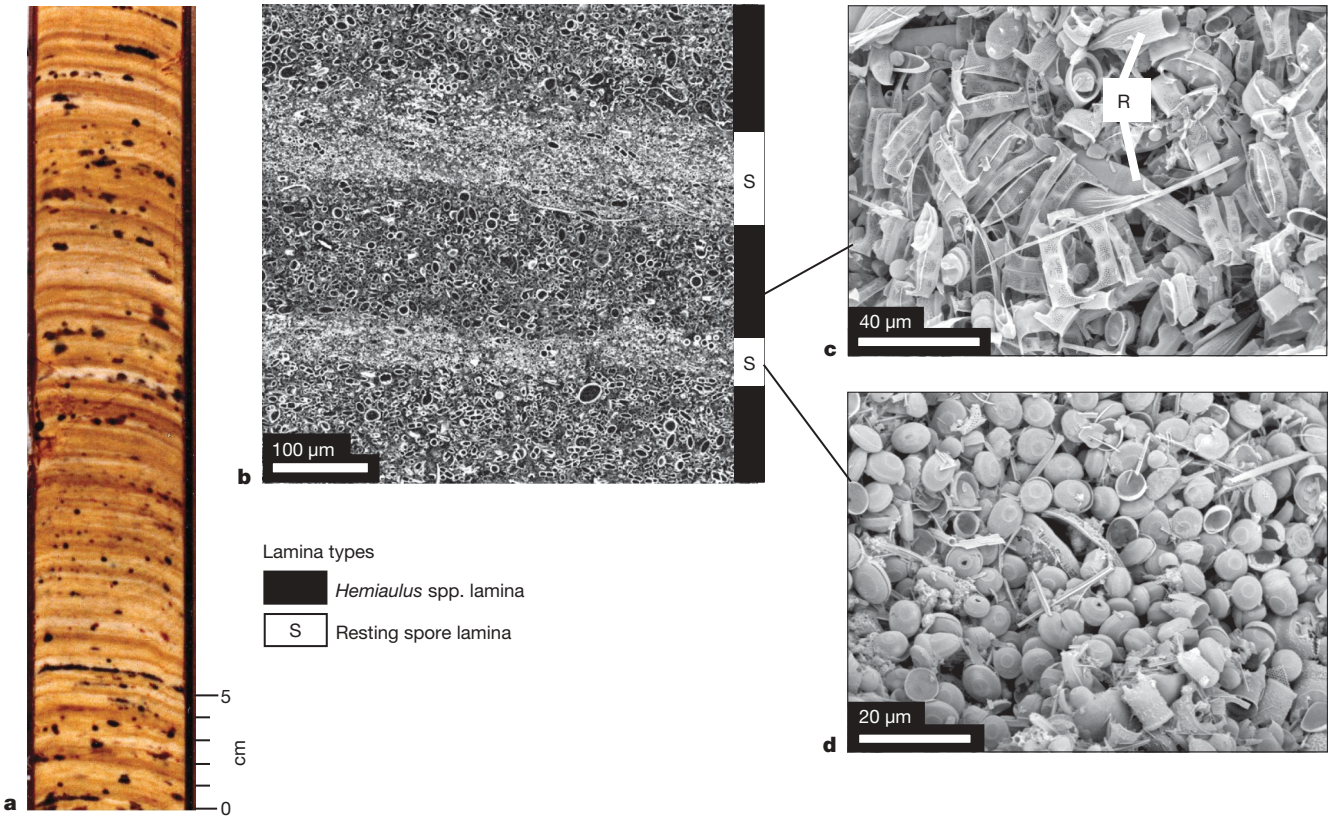


Figure 2 | Composition of the laminated CESAR-6 core. a, Core surface showing the macroscopic colour banding and lamination. Dark blebs are manganese-rich patches[12]. The colour core photo is courtesy of P. Mudie. **b**, BSEI of resin-embedded sediment showing darker, higher porosity, diatom vegetative cell laminae, in which the cross-sections of diatoms are discernible, interbedded with paler, lower-porosity layers composed primarily of diatom resting spores. **c**, **d**, Representative topographic SEM images of diatom vegetative cells, mainly *Hemiaulus* spp., but including some rhizosolenid diatoms (marked R) (**c**), and *Chaetoceros*-type diatom resting spores (**d**).

revised Canadian Arctic diatom zonation[17]. Diatom analysis of core Fl-437 indicates a similar late Campanian age (71–76 Myr ago)[18]. However, because of conflicting palynological evidence and the lack of any comparable well preserved diatomaceous sequences of early Maastrichtian age dated by other fossil groups, an early Maastrichtian age (69–71 Myr ago) cannot be ruled out (see Supplementary Information).

The core surfaces of both CESAR-6 and Fl-437 displayed macroscopic colour lamination (Fig. 2a), on a millimetre-to-centimetre scale, for which two strongly conflicting interpretations have been proposed. A seasonal alternation of diatom vegetative cells and diatom resting spores was proposed for the Fl-437 core on the basis of sampling from macroscopically visible laminations[2,18], whereas analysis of the CESAR-6 core found no differences in diatom content associated with the macroscopic lamination, but attributed the lamination to variations in iron content related to hydrothermal activity[11,12]. To resolve these competing hypotheses we examined samples using electron microscope techniques (see Methods). Analysis, at a spatial resolution that is only possible using back-scattered electron imagery (BSEI) of resin-embedded sediment, reveals that both cores comprise a regular alternation of laminae composed of two distinctly different diatom assemblages (Fig. 2c, d) that are an order of magnitude thinner than those macroscopically visible. Superimposed on these laminae of differing diatom composition are changes in the concentration of fine Fe-rich coatings that define the macroscopically observed lamination or colour banding on the millimetre-to-centimetre scale[12] (Supplementary Fig. S2). Thus, elements of both prior interpretations appear correct. The preservation of undisturbed laminae is consistent with the presence of anoxic bottom waters, as would be expected in a basin with strongly stratified waters, analogous to the modern Black Sea, with an excess of precipitation over evaporation and only shallow water connections to the world ocean through the Turgay and Fram straits and increasingly restricted Western Interior Seaway (Fig. 1)[19]. BSEI analysis of the composition and sequence of the laminae provides new insights into late Cretaceous polar oceanography and climate.

We undertook BSEI analysis of a continuous 58 cm interval from the laminated unit in CESAR-6 (core interval 167.5–225.5 cm; see Supplementary Information), complemented by examination of some discrete 2-cm-long samples from core Fl-437. To verify the integrity of the CESAR-6 section, detailed diatom assemblage counts using optical microscopy were undertaken at 10-cm intervals and these indicate no biostratigraphic change in the diatom flora over the continuously studied interval. More than 3,600 individual laminae of different composition were identified throughout the CESAR-6 interval. We found no evidence for erosional contacts involving truncation of laminae or for current reworking such as ripple cross lamination or lag deposits. Our extensive experience of hemipelagic and pelagic sediments indicates that the observed fabrics are likely to represent an undisturbed record of successive water column flux events.

The diatom laminae within both cores comprise alternations of resting spores with vegetative cells (Fig. 2). The resting spores occur as nearly monospecific laminae of a small hyaline resting spore resembling modern *Chaetoceros* spp. (Fig. 2d). Some laminae of this *Chaetoceros*-type resting spore are mixed with *Hemiaulus tumidicornis* resting spores, and other less-common laminae include *Skeletonema subantarctica*. In the modern ocean, diatom resting spore flux is characteristic of the spring bloom in temperate and high latitudes and is normally related to the depletion of nutrients towards the end of the bloom[20]. *Skeletonema* spp. are also a common component of flux from the spring bloom in areas of good opal

preservation[14]. The characteristic spring flux of resting spores is also typical of Neogene (23 Myr ago) to recent laminated sediments in which the genus *Chaetoceros* is predominant[14]. In the Cretaceous Arctic Ocean (as today) there would have been a rapid transition over about one month from complete darkness to continual daylight, and without permanent ice cover this would have led to the initiation of the spring bloom and to the development of a strong seasonal thermocline (Fig. 3a).

The laminae of diatom vegetative cells that succeed the spring bloom resting spores reflect a different origin. These laminae are dominated by *Hemiaulus* spp. (mainly *H. antiquus* and *H. gleseri* with subordinate *H. danicus* and *H. elegans*) together with *Trochosiropsis polychaeta*, *Anaulus sibiricus* and *Rhizosolenia* spp., all of which may form nearly monospecific (or, with *Hemiaulus*, nearly monogeneric) laminae (Fig. 2c). Recent observations within stratified open-ocean settings such as the North Pacific subtropical gyre, have identified major blooms of *Hemiaulus hauckii* in association with the nitrogen-fixing, intracellular cyanobacterial symbiont, *Richelia intracellularis*, some of which produce significant export[3]. The Pleistocene and Pliocene epochs' organic carbon-rich, Mediterranean sapropels also contain *Hemiaulus hauckii* as a lamina-forming component[4], together with evidence for N_2-fixation[21]. The Cretaceous laminae also contain rhizosolenid diatoms ('R' in Fig. 2c), which are often associated with *Hemiaulus* in the modern ocean and in the Mediterranean sapropels and may not only maintain symbiosis with *R. intracellularis* but may also obtain nitrogen through vertical migration to the nutricline[22]. Rhizosolenid diatoms may also thrive at low light levels in deep chlorophyll maxima[14]. Thus, the diatom vegetative cell laminae of the Arctic Cretaceous are consistent with production within stratified waters (Fig. 3b).

By analogy with the blooms of the North Pacific subtropical gyre, nutrients may have been provided by episodic storm-induced eddies[3]. Such blooms might also have been driven by N_2-fixing intracellular symbionts, given that Si but not N would have been provided by riverine input or through vertical migration of the diatoms to tap deep nutrient sources. Strong seasonal stratification would, indeed, have been expected in the polar summer. Furthermore, Campanian reconstructions from Arctic Canadian Islands indicate significant river run-off to the Arctic Ocean[23] and this freshwater influx would have enhanced stratification as well as introducing nutrients. Subsequently, with the breakdown of the seasonal thermocline in the rapid transition to the polar night, the diatoms adapted to the stratified conditions would have been rapidly sedimented in a 'fall (autumn) dump'[14] (Fig. 3). Pulses of sedimentation may also have occurred during the summer in response to post-bloom nutrient limitation as has been proposed for some *Hemiaulus hauckii* blooms in the North Pacific subtropical gyre[24]. In view of the strong seasonal affinity of the two lamina types (Figs 2 and 3), we can argue that the combined lamina pair represents an annual sediment couplet.

The absence of any fragmentation of the diatoms strongly suggests that a rapid post-bloom flux to the sediments occurred before zooplankton populations could respond, possibly complemented by the accumulation of diatom aggregates in 'grazer-proof' concentrations in deep chlorophyll maxima before a flux event (Fig. 3). This is in contrast to Campanian coastal and shelf sediments from the Canadian Arctic Islands to the south in which there is much fragmented diatom material[23]. The absence of the near-shore, benthic diatom *Paralia* in both the CESAR-6 and Fl-437 cores (present in the Campanian shelf sediments to the south[23]) further indicates a pelagic basinal environment.

Unlike in the modern ocean where the genus *Hemiaulus* is represented by only four species, *Hemiaulus* was one of the dominant genera of the late Cretaceous with over 40 species[25] and may have been better adapted to the less mixed and more stratified seas of the Mesozoic. Intriguingly, biomarker evidence is emerging of cyanobacterial contributions to Cretaceous black shales[26]. It may be that these were produced by cyanobacterial diatom symbionts analogous to *R. intracellularis*, a known symbiont of three of the four extant

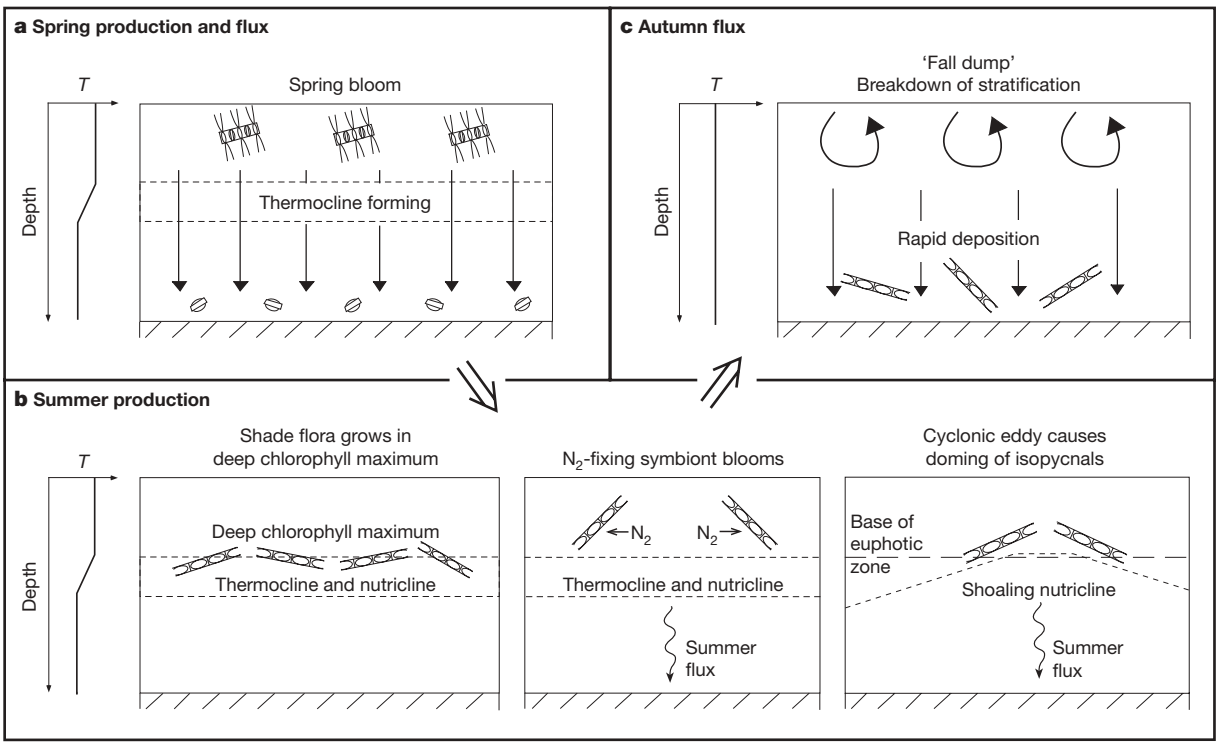


Figure 3 | Seasonal cycle of production and flux in the late Cretaceous Arctic Ocean. a, Spring bloom. After winter mixing has introduced nutrients to the surface, when light returns and at onset of stratification, the spring bloom occurs. When nutrients are consumed, diatom resting spores form and settle rapidly. **b**, Scenarios for summer production (from left to right): accumulation in a deep chlorophyll maximum aided by the ability to grow in low light conditions; symbiosis with nitrogen-fixing cyanobacteria providing nitrate for growth; storm-generated eddies causing doming of isopycnals introducing nutrients to the photic zone. **c**, Autumn flux (the fall dump) when stratification breaks down. *T*, temperature.

Hemiaulus species. Together with complementary biomarker indicators of rhizosolenid diatoms in the late Cretaceous[9], these lines of evidence suggest that diatom algae may already have been key players in the marine biological carbon pump by late Cretaceous times.

An upwelling origin was originally suggested for the diatomaceous sediments of the Alpha ridge[2], but it is clear from our results that much of the production and export was from highly stratified waters with an initial spring bloom episode and subsequent production in the stratified summer surface layer. The diatom sediment concentrations we measured in the Fl-437 and CESAR-6 cores range between 2×10^9 and 8×10^9 valves per gram of dry sediment, and represent exceptional abundances equalling modern values for the most productive areas of the Southern Ocean. This Cretaceous production, dominated by diatoms adapted to stratified conditions of the polar summer, may also be a pointer to future trends in the modern ocean. With increasing CO_2 levels and global warming giving rise to increased ocean stratification[5], this style of production may become increasingly important.

Our evidence for seasonal diatom production and flux testify to an ice-free summer, but the presence of enigmatic lithogenic material suggests the intermittent presence of winter sea ice. Elongate lenses, some isolated grains and very occasional thin laminae of poorly sorted clastic sediment, typically of clay through silt up to fine-sand grade, occur within about 30% of the annual lamina couplets (Fig. 4). The absence of sorting or any other evidence of density current activity suggests a rafting origin (see Supplementary Information). The lack of material coarser than fine sand rules out glacial ice, or sea ice formed from anchor ice flotation, but the size spectrum of particles (in particular the cut-off at fine-sand grade) is typical of that of turbid ice, which is formed in the modern Arctic by sediment entrainment by frazil ice during autumn storms[27,28]. Such early formed ice is unconsolidated and mobile, facilitating long-range ice rafting[27]. Although a 15 °C mean annual sea surface temperature for the late Cretaceous Arctic Ocean has been postulated[9], subsequent papers now propose that the TEX86 method yields a summer sea surface temperature maximum rather than an annual average[29], allowing the possibility of significantly colder winter temperatures. Our evidence for the presence of sea ice would be consistent with indications of possible frosts from Campanian–Maastrichtian age tree rings from Ellesmere Island[6]; estimates of polar temperatures down to –10 °C from late Campanian–Maastrichtian Alaskan vertebrate enamels[7], as well as models of the late Cretaceous Ocean that suggest the formation of seasonal sea ice[8].

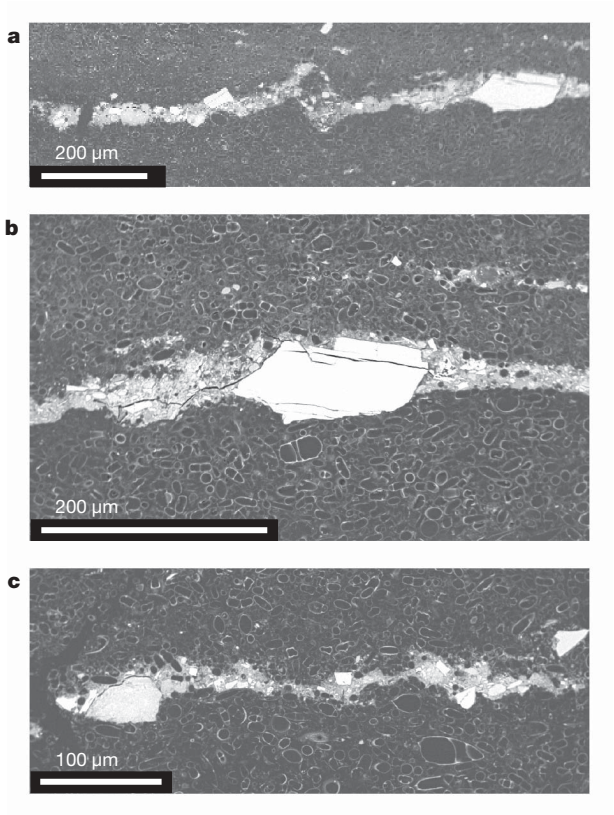


Figure 4 | Evidence for ice rafting. Thin laminae of poorly sorted terrigenous sediment occurring dominantly within the spring diatom bloom layer. **a**, General view of lamina. **b**, Detail of sand-sized grain in **a**. **c**, Lamina containing silt and sand-sized grains within clay.

METHODS SUMMARY

Samples were obtained from the CESAR-6 core using a sediment slab cutter to extract slabs of the still-wet diatom ooze. Subsamples of the cut slabs were embedded with resin using fluid-displacive embedding techniques. Six 2-cm palaeomagnetic sample cubes of already dried sediment from the Fl-437 core were embedded by direct infiltration of resin. Polished thin-sections were prepared and imaged in a scanning electron microscope (SEM) using BSEI to resolve lamina composition. The polished thin-sections used for BSEI were also imaged with a flat-bed scanner at a resolution of 2,400 dots per inch, to examine closely the colour laminations previously reported by visual examination of the core and to compare with the results of BSEI analysis. To complement BSEI, counterpart samples of the sectioned material were prepared for topographic SEM imaging to enable microfossil identification. Detailed diatom micropalaeontology was undertaken using conventional quantitative diatom abundance preparation techniques and optical microscopy. The analysis of a series of diatom micropalaeontology samples from the CESAR-6 core was used to assess the biostratigraphic continuity of the laminated interval.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 September 2008; accepted 13 May 2009.

1. Graversen, R. G., Mauritsen, T., Tjernstrom, M., Kallen, E. & Svensson, G. Vertical structure of recent Arctic warming. *Nature* **451,** 53–56 (2008).
2. Kitchell, J. A. & Clark, D. L. Late Cretaceous-Paleogene paleogeography and paleocirculation: evidence of north polar upwelling. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **40,** 135–165 (1982).
3. Dore, J. E., Letelier, R. M., Church, M. J., Lukas, R. & Karl, D. M. Summer phytoplankton blooms in the oligotrophic North Pacific Subtropical Gyre: historical perspective and recent observations. *Prog. Oceanogr.* **76,** 2–38 (2008).
4. Kemp, A. E. S., Pearce, R. B., Koizumi, I., Pike, J. & Rance, S. J. The role of mat-forming diatoms in the formation of Mediterranean sapropels. *Nature* **398,** 57–61 (1999).
5. Sarmiento, J. L., Hughes, T. M. C., Stouffer, R. J. & Manabe, S. Simulated response of the ocean carbon cycle to anthropogenic climate warming. *Nature* **393,** 245–249 (1998).
6. Falcon-Lang, H. J., MacRae, R. A. & Csank, A. Z. Palaeoecology of late Cretaceous polar vegetation preserved in the Hansen Point volcanics, NW Ellesmere Island, Canada. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **212,** 45–64 (2004).
7. Amiot, R. *et al.* Latitudinal temperature gradient during the Cretaceous Upper Campanian–Middle Maastrichtian: $\delta^{18}O$ record of continental vertebrates. *Earth Planet. Sci. Lett.* **226,** 255–272 (2004).
8. Otto-Bliesner, B. L., Brady, E. C. & Shields, C. Late Cretaceous ocean: coupled simulations with the National Center for Atmospheric Research Climate System Model. *J. Geophys. Res.* **107,** doi:10.1029/2001JD000821 (2002).
9. Jenkyns, H. C., Forster, A., Schouten, S. & Damste, J. S. S. High temperatures in the Late Cretaceous Arctic Ocean. *Nature* **432,** 888–892 (2004).
10. Moran, K. *et al.* The Cenozoic palaeoenvironment of the Arctic Ocean. *Nature* **441,** 601–605 (2006).
11. Mudie, P. J., Stoffyn-Egli, P. & Van Wagoner, N. A. Geological constraints for tectonic models of the Alpha Ridge. *J. Geodynam.* **6,** 215–236 (1986).
12. Stoffyn-Egli, P. Iron and manganese micro-precipitates within a Cretaceous biosiliceous ooze from the Arctic Ocean: possible hydrothermal source. *Geo-Mar. Lett.* **7,** 223–231 (1987).
13. Sarthou, G., Timmermans, K. R., Blain, S. & Treguer, P. Growth physiology and fate of diatoms in the ocean: a review. *J. Sea Res.* **53,** 25–42 (2005).
14. Kemp, A. E. S., Pike, J., Pearce, R. B. & Lange, C. B. The "Fall dump"—a new perspective on the role of a "shade flora" in the annual cycle of diatom production and export flux. *Deep-Sea Res. II* **47,** 2129–2154 (2000).
15. Harwood, D. M., Nikolaev, V. A. & Winter, D. M. in *Pond Scum to Carbon Sink: Geological and Environmental Applications of the Diatoms* (ed. Starratt, S.) 33–59 (Paleontological Society, 2007).
16. Wagner, T., Damste, J. S. S., Hofmann, P. & Beckmann, B. Euxinia and primary production in Late Cretaceous eastern equatorial Atlantic surface waters fostered

orbitally driven formation of marine black shales. *Paleoceanography* **19,** doi:10.1029/2003PA000898 (2004).
17. Tapia, P. M. & Harwood, D. M. Upper Cretaceous diatom biostratigraphy of the Arctic archipelago and northern continental margin, Canada. *Micropaleontology* **48,** 303–342 (2002).
18. Dell'Agnese, D. J. & Clark, D. L. Siliceous microfossils from the warm Late Cretaceous and Early Cenozoic Arctic Ocean. *J. Paleontol.* **68,** 31–47 (1994).
19. Hay, W. W. *et al.* in *Evolution of the Cretaceous Ocean-Climate System* (eds Barrera, E. & Johnson, C. C.) 1–47 (Geological Society of America Special Paper 332, 1999).
20. McQuoid, M. R. & Hobson, L. A. Importance of resting stages in diatom seasonal succession. *J. Phycol.* **31,** 44–50 (1995).
21. Sachs, J. P. & Repeta, D. J. Oligotrophy and nitrogen fixation during eastern Mediterranean sapropel events. *Science* **286,** 2485–2488 (1999).
22. Singler, H. R. & Villareal, T. A. Nitrogen inputs into the euphotic zone by vertically migrating Rhizosolenia mats. *J. Plankton Res.* **27,** 545–556 (2005).
23. Chin, K. *et al.* Life in a temperate polar sea: a unique taphonomic window on the structure of a Late Cretaceous Arctic marine ecosystem. *Proc. R. Soc. B* **275,** 2675–2685 (2008).
24. Scharek, R., Tupas, L. M. & Karl, D. M. Diatom fluxes to the deep sea in the oligotrophic North Pacific gyre at Station ALOHA. *Mar. Ecol. Prog. Ser.* **182,** 55–67 (1999).
25. Harwood, D. M. & Nikolaev, V. A. in *Siliceous Microfossils* (eds Blome, C. D. *et al.*) 81–106 (Paleontological Society, 1995).
26. Kuypers, M. M. M., van Breugel, Y., Schouten, S., Erba, E. & Damste, J. S. S. N_2-fixing cyanobacteria supplied nutrient N for Cretaceous oceanic anoxic events. *Geology* **32,** 853–856 (2004).
27. Kempema, E. W., Reimnitz, E. & Barnes, P. W. Sea ice sediment entrainment and rafting in the Arctic. *J. Sedim. Petrol.* **59,** 308–317 (1989).
28. Smedsrud, L. H. Frazil-ice entrainment of sediment: large-tank laboratory experiments. *J. Glaciol.* **47,** 461–471 (2001).
29. Sluijs, A. *et al.* Subtropical Arctic ocean temperatures during the Palaeocene/Eocene thermal maximum. *Nature* **441,** 610–613 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The research was supported by the award of a NERC Research Studentship (A.D.) and a NERC Research Grant (A.E.S.K. and J.P.). We are grateful to P. Mudie for facilitating sampling of the CESAR-6 core and to D. Clark for providing samples from core FL-437. We thank R. Pearce for assistance with electron microscopy and K. Davis for expertise in drafting. We also thank P. Wilson and H. Pälike for comments on the manuscript.

Author Contributions Preliminary SEM and diatom studies of the material were undertaken by J.P. and A.E.S.K. Detailed SEM lamina studies and diatom analysis were performed by A.D. during PhD studies supervised by A.E.S.K. and J.P. All authors contributed to interpretation. The manuscript was written by A.E.S.K. and incorporates comments from all others.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.E.S.K. (aesk@noc.soton.ac.uk).

METHODS

Sampling and sample preparation. Samples were obtained from the CESAR-6 core using a sediment slab cutter and sampling protocols following standard procedures[30]. Subsamples of these were embedded in epoxy resin using fluid-displacing techniques (CESAR-6)[30]. Thin-sections of embedded sediment were cut with boundaries oblique to the lamination to ensure continuity between thin-sections, which were then polished and carbon-coated. Six discrete samples of sediment from core Fl-437 were provided in 2-cm palaeomagnetic plastic sample cubes. These samples that were already dried were embedded by direct infiltration of low-viscosity epoxy resin.

Scanning electron microscope methods. Overlapping polished thin-sections were made for the CESAR-6 interval spanning 138–305 cm and analysed using a high-vacuum LEO 1450VP SEM. Low-resolution (66×) BSEI photo-mosaic base maps were produced for each polished thin-section by taking 10% overlapping images. Higher resolution (750× and 1,000×) photo-mosaics were then produced for more detailed analysis of the interval 167.5–222.5 cm. Measurements of individual lamina thickness were made from the high-resolution photo-mosaics using the methods of Francus[31]. This method involves the use of the Adobe Photoshop program to mark the boundaries between individual laminae on a digitized photo-mosaic of the polished section. The resultant 'path' is then exported to Microsoft Excel, where a macro converts the measurements into actual thickness using the known quantities of the number of pixels per centimetre on the image (obtained from Photoshop) and the size of individual pixels (obtained from the SEM).

To help us constrain the species make-up of individual laminae, the off-cut portions of individual polished thin-sections were kept and used to make topographic stubs. Laminae of interest were carefully fractured and fixed onto SEM stubs using carbon-based wax for SEM analysis. As a complement to the topographic stubs, peel slides were taken of the raw material using double-sided carbon tape. These provided topographic images of the diatoms *in situ* with very little disturbance to the sedimentary microfabric. As microfabrics were analysed primarily from the cross-sectional images obtained from the polished thin-sections, the peels greatly facilitated the task of diatom identification within individual laminae. Suitably sized pieces of tape were cut to fit onto standard SEM glass slides. The tape was placed on the surface of the sediment block and gently pressed to ensure contact was made along the entire length of the tape. The tape was then carefully peeled off and affixed onto a glass slide. Two peels were made for each section studied and affixed to the same glass slide. The slides were subsequently coated in gold in preparation for use with the SEM. The polished thin-sections used for BSEI were also imaged with a flatbed scanner at a resolution of 2,400 dots per inch, to examine closely the colour laminations previously reported by visual examination of the core and to compare with the results of BSEI analysis.

Diatom micropalaeontology. Diatom strew slides were prepared from 19 samples of CESAR-6 Unit 4 (see Supplementary Information). To ensure several laminae were included in all slides, samples comprised material homogenized from 1 cm of the core. Samples were taken at intervals of 10 cm, from 138–298 cm, with additional samples taken at 134 and 305 cm. Samples were digested and prepared using an adaptation of the method of Scherer[32] (Claire S. Allen, personal communication, 2004). This method was found to produce an even spread of diatomaceous material over the whole coverslip, without bunching of frustules. Quantitative analyses were performed at magnifications of 500×, with identifications checked at 1,000×. Counts were made along several traverses of the cover slips.

Quantitative diatom abundances, or concentrations, were also calculated for two samples from the CESAR-6 core (210.5–210.8 cm and 211.8–212.1 cm), and one from the Fl-437 core (sample Fl-437-13-2), prepared using the digestion and random settling method of Bodén[33]. Diatoms were counted using 1,000× magnification and concentrations (in valves per gram)were calculated using the following formula:

$$\text{Concentration} = (\text{WD} \times \text{AC} \times \text{ND})/(\text{WS} \times \text{AV} \times \text{NV} \times \text{VS})$$

where WD is the mass of sample solution in grams, AC is the area of the settling container in mm^2, ND is the number of diatom valves counted, WS is the dry weight of the initial sample in grams, AV is the area of the field of view (or transect) in mm^2, NV is the number of fields of view (or transects) and VS is the mass of the subsample solution in grams[33].

30. Kemp, A. E. S., Dean, J., Pearce, R. B. & Pike, J. in *Tracking Environmental Change Using Lake Sediments* Vol. 2 *Physical and Geochemical Methods* (eds Last, W. M. & Smol, J. P.) 7–22 (Kluwer, 2001).
31. Francus, P., Keimig, F. & Besonen, M. An algorithm to aid varve counting and measurement from thin-sections. *J. Paleolimnol.* **28,** 283–286 (2002).
32. Scherer, R. P. A new method for the determination of absolute abundance of diatoms and other silt-sized sedimentary particles. *J. Paleolimnol.* **12,** 171–179 (1995).
33. Bodén, P. Reproducibility in the random settling method for quantitative diatom analysis. *Micropalaeontology* **37,** 313–319 (1991).

LETTERS

Bone-marrow adipocytes as negative regulators of the haematopoietic microenvironment

Olaia Naveiras[1], Valentina Nardi[1]*, Pamela L. Wenzel[1]*, Peter V. Hauschka[2], Frederic Fahey[3] & George Q. Daley[1]

Osteoblasts and endothelium constitute functional niches that support haematopoietic stem cells in mammalian bone marrow[1–3]. Adult bone marrow also contains adipocytes, the number of which correlates inversely with the haematopoietic activity of the marrow. Fatty infiltration of haematopoietic red marrow follows irradiation or chemotherapy and is a diagnostic feature in biopsies from patients with marrow aplasia[4]. To explore whether adipocytes influence haematopoiesis or simply fill marrow space, we compared the haematopoietic activity of distinct regions of the mouse skeleton that differ in adiposity. Here we show, by flow cytometry, colony-forming activity and competitive repopulation assay, that haematopoietic stem cells and short-term progenitors are reduced in frequency in the adipocyte-rich vertebrae of the mouse tail relative to the adipocyte-free vertebrae of the thorax. In lipoatrophic A-ZIP/F1 'fatless' mice, which are genetically incapable of forming adipocytes[5], and in mice treated with the peroxisome proliferator-activated receptor-γ inhibitor bisphenol A diglycidyl ether, which inhibits adipogenesis[6], marrow engraftment after irradiation is accelerated relative to wild-type or untreated mice. These data implicate adipocytes as predominantly negative regulators of the bone-marrow microenvironment, and indicate that antagonizing marrow adipogenesis may enhance haematopoietic recovery in clinical bone-marrow transplantation.

At birth, haematopoietic red marrow occupies virtually the entirety of the bone-marrow space, but with age, non-haematopoietic fatty marrow gradually predominates[7]. This 'fatty degeneration' of the marrow is a dynamic and reversible process[8,9]. To investigate whether marrow adipocytes influence haematopoiesis, we surveyed the mouse skeleton for bones that harbour fatty marrow under normal conditions. We found that the spine of adult mice manifests a proximal to distal gradient of bone-marrow adipocytes: thoracic vertebrae are virtually adipocyte-free, whereas vertebrae starting at the level of the third or fourth tail segments are highly adipocytic (Fig. 1a). We isolated bone marrow from the thoracic and tail vertebrae of 12-week-old mice and quantified the haematopoietic stem-cell and short-term progenitor compartments both phenotypically and functionally (schema in Supplementary Fig. 1). Bone marrow from tail vertebrae contains only 25% as many $CD45^+$ haematopoietic cells per segment as thoracic bone marrow, thus confirming the reduced overall haematopoietic cellularity evident by histology (Fig. 1b and Supplementary Fig. 2a). Using flow cytometry, we determined the relative frequency of haematopoietic stem cells (HSCs; $cKit^+Lin^-Sca1^+Flk2^-$ (KLSF)), multipotent progenitors (MPPs), common myeloid progenitors (CMPs), granulocyte–macrophage progenitors (GMPs) and megakaryocyte–erythroid progenitors (MEPs) in these different regions of the spine[10]. The percentage of all progenitor classes was reduced 2–3-fold in the $CD45^+$ haematopoietic cells of adipocyte-rich bone marrow of the tail vertebrae compared with non-adipocytic bone marrow from the thoracic vertebrae (Fig. 1c and Supplementary Fig. 2b). Congruent with the fluorescence-activated cell sorting (FACS) phenotype data, long-term repopulating HSCs, short-term repopulating progenitors, spleen colonies and methylcellulose colony-forming units (c.f.u.) were reduced 1.5–3-fold in adipocyte-rich bone marrow from tail vertebrae as compared with adipocyte-free bone marrow from the thoracic vertebrae (Fig. 1d–f and Supplementary Fig. 2c). These phenomena are neither due to age nor weight-bearing status, for we observed a similar reduction in the frequency of short-term progenitors in younger (4 week-old) and older mice (13 months), as well as in another consistently fatty yet weight-bearing region of the mouse skeleton, the distal tibia (Supplementary Fig. 3a–f). Reduced frequency of CMPs and primitive c.f.u. also accompany the increased bone-marrow adiposity of femurs from leptin-deficient obese mice (Supplementary Fig. 4a–d). Moreover, we found that bone-marrow-derived adipocytes reduce the expansion of haematopoietic cells in stromal transwell co-cultures, indicating that adipocytes release diffusible inhibitors of haematopoiesis (Supplementary Fig. 5a–d). We therefore conclude that adipocyte-rich marrow is associated with lower absolute levels and relative numbers of haematopoietic progenitors.

To investigate the mechanism of reduced haematopoietic activity of adipocytic bone marrow, we performed cell-cycle analysis of the progenitor compartments. In all cases, we found fewer progenitors in the replicating phases of the cell cycle (S/G2/M) within the adipocyte-rich bone marrow (Fig. 1g and Supplementary Fig. 6a). Early progenitors (HSC/MPP) presented no significant difference in their G0/G1 ratio, whereas late progenitors (GMP/CMP/MEP) presented a significant increase in the G0/G1 ratio (Supplementary Fig. 6b). To determine if the slow-cycling HSCs within the tail bone marrow were functional, as opposed to senescent, we FACS-sorted HSCs (KLSF) and transplanted them competitively into lethally irradiated mice. We observed no difference in repopulating activity between HSCs from tail and thorax in the first month after transplant. However, multilineage long-term engraftment was significantly higher in HSCs purified from tail bone marrow (Fig. 1h), indicating that the slow-cycling progenitors in adipocytic tail bone marrow are relatively quiescent and not senescent. This is consistent with the relative predominance of $CD34^{low}$ HSCs within the KLSF fraction of tail bone marrow (Fig. 1i), a phenotype associated with long-term repopulation activity[11]. Taken together, our data establish tail vertebrae as a model for the study of fatty marrow in the mouse, and demonstrate

[1]Division of Pediatric Hematology/Oncology, Children's Hospital Boston and Dana Farber Cancer Institute; Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School; Division of Hematology, Brigham and Women's Hospital; Harvard Stem Cell Institute; Manton Center for Orphan Diseases; Howard Hughes Medical Institute, Boston, Massachusetts 02115, USA. [2]Departments of Orthopaedic Surgery and Oral and Developmental Biology, Harvard Medical School and School of Dental Medicine, Boston, Massachusetts, 02115, USA. [3]Department of Radiology, Division of Nuclear Medicine/PET, Children's Hospital Boston, Harvard Medical School, 300 Longwood Avenue, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

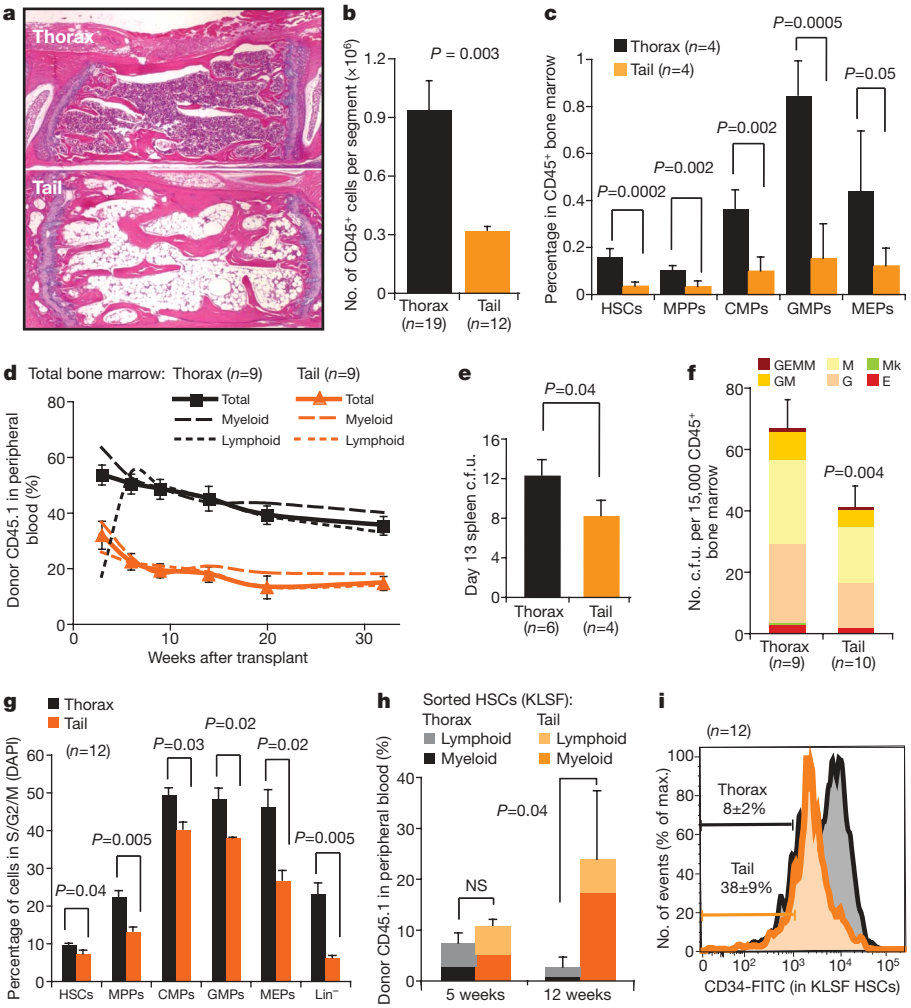


Figure 1 | Haematopoietic stem cells and progenitors are reduced in number, frequency and cycling capacity in adipocyte-rich bone marrow during homeostasis. **a**, Haematoxylin and eosin stain of decalcified thoracic vertebra (top) and fourth tail segment (bottom) of 12-week-old C57BL/6J mice. **b**, Absolute number of haematopoietic cells ($CD45^+$) per vertebral segment. **c**, Progenitor frequency within the haematopoietic compartment ($CD45^+$). **d**, Competitive engraftment (250,000 CD45.1 tail or thorax bone-marrow cells; 250,000 CD45.2 competitor bone-marrow cells). **e**, **f**, Day 13 spleen colony assay (**e**) and c.f.u. progenitor assay (**f**) from tail and thorax bone marrow. E, erythrocyte; G, granulocyte; GEMM, granulocyte–erythrocyte–megakaryocyte–macrophage; GM, granulocyte–macrophage; M, macrophage; Mk, megakaryocyte. **g**, Progenitor cell-cycle analysis. The average percentage of cells in S/G2/M transition is shown (±s.e.m.). **h**, One-hundred tail and thorax bone marrow sorted HSCs ($cKit^+Lin^-Sca1^+Flk2^-$; >95% purity) were transplanted competitively, then engraftment in peripheral blood was monitored. **i**, CD34 expression within the HSC fraction (KLSF, $cKit^+Lin^-Sca1^+Flk2^-$); the percentage of $CD34^{low}$ cells within the KLSF fraction is indicated. All error bars indicate s.e.m.

that adipocyte-rich marrow manifests altered haematopoiesis. HSCs and short-term progenitors are functionally reduced on a per cell basis in fatty marrow owing to reduced cycling at the level of the HSC, MPP and CMP compartments.

To determine whether the association between adipocytic marrow and reduced haematopoietic progenitor frequency is purely correlative, or whether adipocytes actively compromise haematopoiesis, we studied the lipoatrophic 'fatless' A-ZIP/F1 mouse, which cannot form adipocytes owing to the expression of a dominant-negative form of C/EBP under the adipocyte fatty-acid-binding protein 4 (Fabp4, also called aP2) promoter[5]. In contrast to wild-type mice, we found that the absence of adipocytes in fatless A-ZIP/F1 mice rescued haematopoiesis in the tail, such that A-ZIP/F1 mice presented no significant difference in the frequency of c.f.u. from thorax or tail bone marrow (Supplementary Fig. 7a, b), indicating that compromised osteogenesis due to the non-weight-bearing nature of these bones cannot explain the haematopoietic defect of wild-type, adipocyte-rich tail vertebrae[12]. Notably, although fatless A-ZIP/F1 mice are diabetic, their blood counts were similar to controls during homeostasis, and their femoral bone marrow showed no competitive advantage over bone marrow from wild-type littermates, arguing that the diabetic environment does not account for the observed alterations in the haematopoietic compartment (Supplementary Fig. 8a–c). We therefore conclude that the presence of adipocytes is necessary to observe reduced haematopoiesis in adipocyte-rich tail bone marrow.

We then analysed the effect of adipocytes on haematopoietic recovery after bone-marrow transplantation. Between the second and fourth week after lethal irradiation, the bone-marrow space throughout the mouse skeleton is replaced by adipocytes. During this post-transplant period mice (and human patients) depend on fast-cycling, short-term haematopoietic progenitors to rescue their otherwise lethal pancytopaenia[13]. Given our previous data, we predicted that the compromised adipogenesis in the A-ZIP/F1 mouse would enhance haematopoietic recovery in the post-transplant period through increased proliferation of short-term progenitors. We transplanted wild-type bone marrow (CD45.2) into either wild-type or fatless A-ZIP/F1 littermates (CD45.1, Fig. 2a). In contrast to control mice, A-ZIP/F1 fatless mice exposed to lethal doses of irradiation produced markedly fewer adipocytes in the marrow

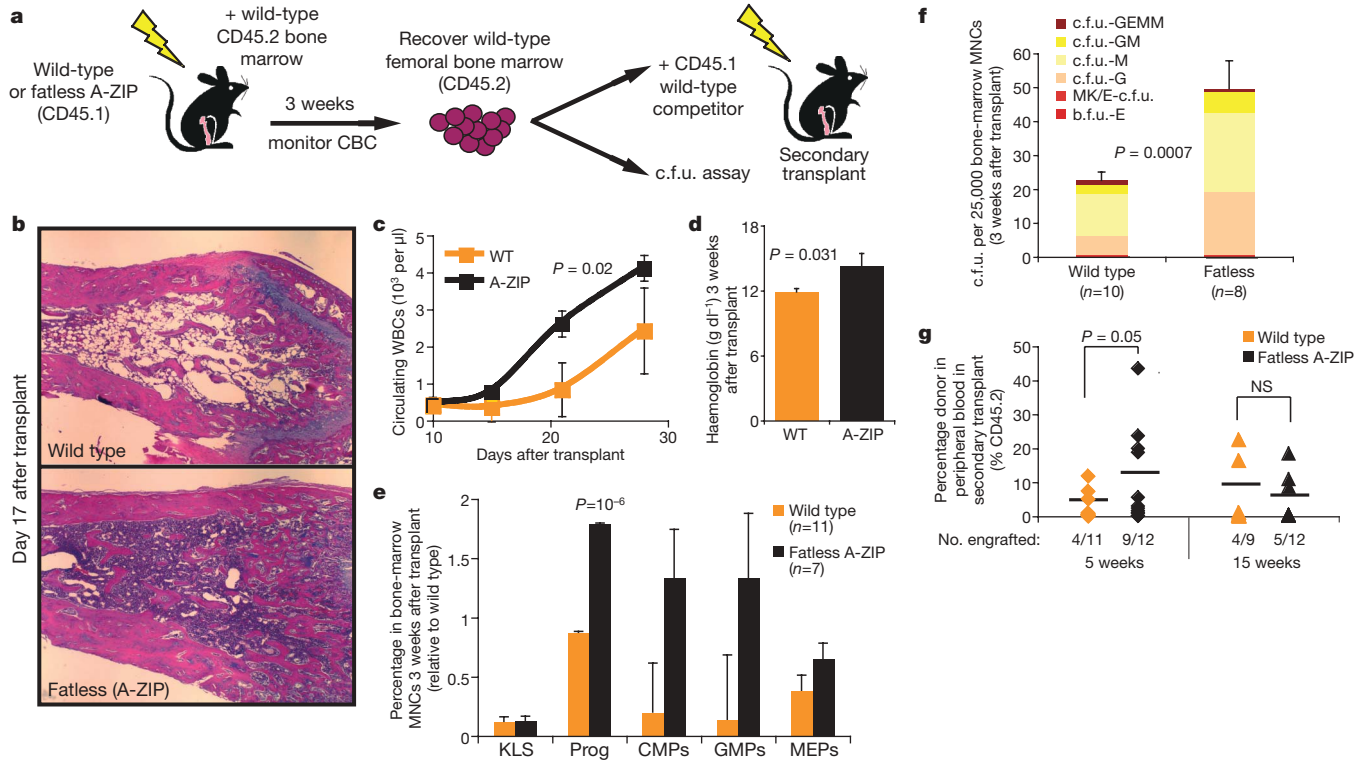


Figure 2 | Lack of bone-marrow adipocytes after irradiation in fatless mice enhances haematopoietic progenitor expansion and post-transplant recovery. **a**, Experimental design. Wild-type FVB or fatless FVB.A-ZIP/F 16-week-old mice (CD45.1) were lethally irradiated and transplanted with 200,000 CD45.2, MHC-compatible DBA/1 wild-type bone-marrow cells. Femurs were isolated on day 17–20 after transplant and donor DBA CD45.2 wild-type bone marrow was recovered by high-purity FACS, then used for progenitor assays or competitive serial transplantation. **b**, Femoral haematoxylin and eosin stain in the third week after transplant. Original magnification, ×40. **c**, **d**, White blood cell (WBC) counts (**c**) and haemoglobin levels (**d**) in peripheral blood after primary transplant (±s.e.m.). **e**–**g**, Bone marrow recovered from primary transplants was assayed for relative frequency of progenitors (Prog) by FACS (**e**; ±s.d.), c.f.u. assay (**f**; ±s.e.m.) and secondary competitive transplantation into wild-type recipients (**g**). b.f.u., burst-forming unit; KLS, $cKit^+Lin^-Sca1^+$; Prog, $cKit^+Lin^-Sca1^-$; MNC, mononuclear cell.

cavity (Fig. 2b). We monitored leukocyte recovery in the post-transplant period and found that recovering A-ZIP/F1 fatless mice have up to four times higher leukocyte counts in peripheral blood relative to their wild-type controls (Fig. 2c). We also observed significantly accelerated recovery in the haemoglobin content of peripheral blood (Fig. 2d). Importantly, both wild-type and A-ZIP/F1 fatless recipients showed comparable high-level long-term donor chimaerism after the primary transplant (Supplementary Fig. 8d). In the third week after transplant, we recovered the donor CD45.2 bone marrow from the adipocyte-rich wild-type or the adipocyte-free AZIP/F1 femurs. We found a pronounced increase in haematopoietic progenitors in the recovering CD45.2 bone marrow isolated from fatless A-ZIP/F1 mice as determined by flow cytometry (Fig. 2e), methylcellulose colony forming assays (Fig. 2f and Supplementary Fig. 8e) and short-term competitive repopulation into secondary recipients (Fig. 2g). Collectively, these results indicate that the lack of adipogenesis in A-ZIP/F1 recipient mice enhances haematopoietic recovery after lethal irradiation by enhancing engraftment of short-term progenitors, and further supports the conclusion that adipocytes in fatty marrow hinder haematopoietic progenitor expansion.

During our studies, we observed that bone-marrow ablation in lethally irradiated A-ZIP/F1 fatless mice was accompanied by marked osteogenesis. Trabecular bone was increased in the femurs of transplanted A-ZIP/F1 fatless mice compared to wild-type controls (Fig. 3a), a phenomenon also apparent in the tail and in mice that were lethally ablated but received no transplant (Supplementary Fig. 9a–c). High-resolution micro-computerized tomography (mCT) confirmed a fivefold increase in trabecular bone that was specific to fatless A-ZIP/F1 mice after bone-marrow transplantation (Fig. 3a–c). Incorporation of ^{18}F measured by micro-positron emission tomography (mPET) confirmed an increased bone metabolism, indicating new bone deposition in tails and tibias after bone marrow transplantation that was maximal in the second week after transplant (Fig. 3d, e). Our data show that simultaneous ablation of the haematopoietic and bone-marrow adipocyte compartment can induce osteogenesis, which, as shown by others, promotes a more supportive environment for haematopoietic reconstitution that could explain the positive effect of adipocyte ablation in bone-marrow engraftment[1,2]. Our observation is compatible with a previous report that surgical removal of the fatty marrow in rabbit tibias induces transient haematopoietic infiltration and new osteoid and trabecular bone formation[14]. In addition to creating an osteogenic environment, fatless A-ZIP/F1 mice may accumulate mesenchymal elements that support haematopoietic recovery, or may be deficient in osteoclast elements that would antagonize trabecular bone growth during the recovery phase of lethal irradiation. Importantly, preventing the formation of bone-marrow adipocytes alone does not cause osteogenesis[15], indicating that osteogenesis requires simultaneous ablation of both the adipocytic and haematopoietic compartments. These data suggest a three-way co-regulation of haematopoiesis, osteogenesis and adipogenesis within the bone-marrow compartment.

Finally, we tested whether blocking adipogenesis pharmacologically could enhance bone-marrow engraftment in wild-type mice. The peroxisome proliferator-activated receptor-γ (PPAR-γ) inhibitor bisphenol A diglycidyl ether (BADGE) has been shown to prevent bone marrow adipocyte formation *in vitro* and *in vivo* in models of streptozotocin-induced diabetes[6,15]. Importantly, BADGE does not enhance haematopoietic colony formation *in vitro*, when bone-marrow cells are isolated from their stromal microenvironment (Supplementary Fig. 10). We administered BADGE to lethally

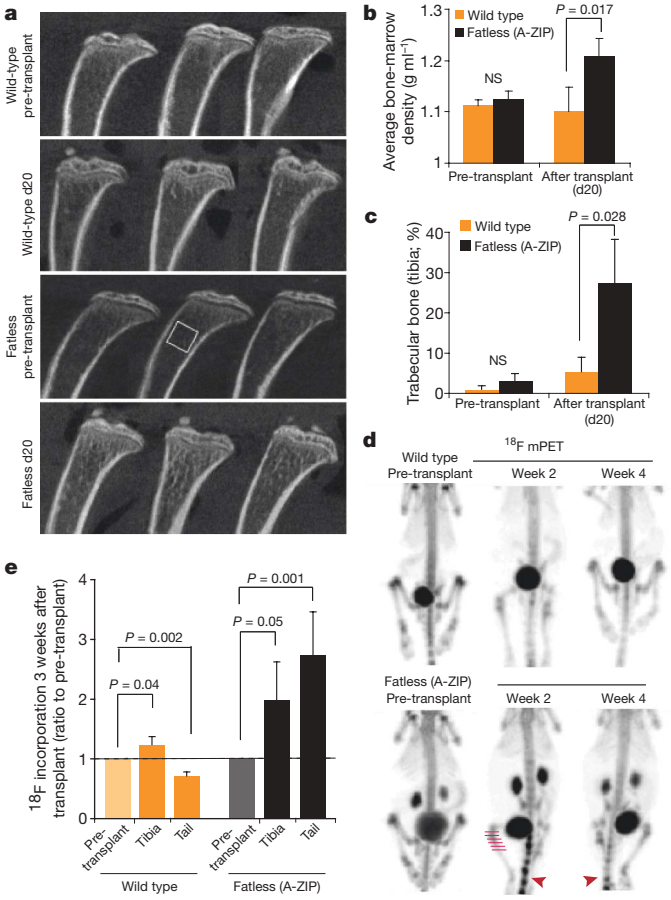


Figure 3 | Ablation of the haematopoietic compartment in fatless A-ZIP/F1 mice during bone-marrow transplantation induces osteogenesis. Analysis of mice transplanted as in Fig. 2. **a**, High-resolution mCT analysis of pre/post-transplant tibias from wild-type or fatless A-ZIP/F1 mice 20 days after lethal ablation. **b**, Average trabecular bone density from panel **a** normalized to a density standard (phantom). **c**, Percentage of bone-marrow space occupied by trabecular bone 20 days after transplantation. **d**, mPET analysis pre/post-transplant. Representative mice shown at three different time points (3–4 analysed per group). Dark areas indicate ^{18}F-labelled NaF uptake in regions of active bone deposition (red arrowheads) and accumulation in urinary apparatus upon excretion. **e**, Quantification of mean ^{18}F-labelled NaF uptake in tibias and proximal tails pre/post-transplantation. Square and lines over micrographs in **a** and **d** indicate quantification regions (see Methods). All error bars indicate s.e.m.

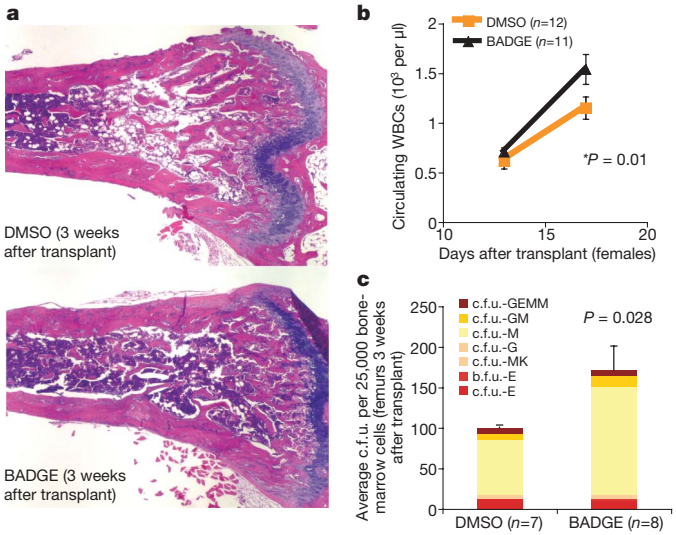


Figure 4 | Pharmacological inhibition of adipocyte formation enhances bone-marrow engraftment in wild-type mice. Bone-marrow transplants were performed in wild-type female FVB mice as described for Fig. 2 except that 30 mg kg^{-1} BADGE or control vehicle (DMSO 10%) was administered through daily intraperitoneal injections from the day before irradiation until day 14 after transplant. **a**, Haematoxylin and eosin stain of femurs from mice killed on day 17 after transplant, when the donor CD45.2 wild-type bone marrow was recovered and purified by FACS. **b**, White blood cell (WBC) counts in peripheral blood on the post-transplant period show accelerated recovery in BADGE-treated mice. **c**, Colony-forming unit (c.f.u.) assay from the recovered donor bone marrow. **b**, **c**, Error bars are s.e.m.

irradiated mice for the 2 weeks after bone-marrow transplantation, and observed successful inhibition of bone-marrow adipocyte formation (Fig. 4a), higher peripheral blood leukocyte counts (Fig. 4b) and an enrichment in c.f.u. (Fig. 4c). Our results demonstrate that the negative influence of adipocytes on post-transplant haematopoietic engraftment can be overcome pharmacologically, and suggest that PPAR-γ inhibitors, or other adipocyte inhibitors such as the novel Fabp4 inhibitor BMS309403 (ref. 16), might serve as adjuvants to haematopoietic recovery in clinical bone-marrow transplantation.

Collectively, our results contradict the classical dogma that adipocytes act as passive space fillers in the marrow. We demonstrate that adipocyte-rich marrow harbours a decreased frequency of progenitors and relatively quiescent stem cells. Moreover, we observe that mice that are genetically deficient in adipogenesis show accelerated haematopoietic recovery after bone-marrow ablation, a phenomenon that can be reproduced pharmacologically in wild-type mice through PPAR-γ inhibition. These results suggest a novel therapeutic approach to enhance haematopoietic engraftment after marrow or cord blood transplantation, or to ameliorate aplasia in genetic bone-marrow failure syndromes. Furthermore, our results suggest a plausible mechanism for the reports of myelosuppression[17–19] in patients treated with the PPAR-γ agonist rosiglitazone, a diabetes drug known to increase marrow adiposity[20].

Our data indicate a predominantly suppressive influence of adipocytes on haematopoiesis within the bone-marrow microenvironment. Bone-marrow adipocytes are less supportive of haematopoiesis *in vitro* than their undifferentiated stromal or pre-adipocytic counterparts, in part due to reduced production of growth factors such as granulocyte–macrophage colony-stimulating factor (GM-CSF) and granulocyte colony-stimulating factor (G-CSF)[21,22]. Moreover, adipose tissue secretes neuropillin-1 (ref. 23), lipocalin 2 (refs 24, 25), adiponectin[26] and TNF-α[27,28], each of which can impair haematopoietic proliferation. Of note, TNF-α and adiponectin inhibit progenitor activity while positively influencing the most primitive HSCs[27,29], suggesting that adipocytes prevent haematopoietic progenitor expansion while preserving the haematopoietic stem-cell pool. Adipocytes and osteoblasts originate from mesenchymal stem cells within the bone marrow, where both compartments hold a reciprocal relationship[30]. Balancing the supportive role of the osteoblast in the HSC niche, our data implicate adipocytes as negative regulators of haematopoiesis. Further studies will address the molecular players involved in the haematopoietic inhibition imposed by fatty marrow.

METHODS SUMMARY

Thorax, tibias and tails were isolated from wild-type C57BL6/J mice. B6.SJL-Ptprca Pep3b/BoyJ mice were used as bone-marrow donors to exploit the CD45.1/CD45.2 allelic system (Jackson Laboratories 0002014). Bone-marrow transplantation and c.f.u. spleen assays were performed on lethally irradiated mice (two 5.5-Gy doses separated by 3 h) and cells were administered by tail vein injection within 24 h of lethal irradiation. For competitive transplantation, samples were competed against 250,000 recipient-matched competitor femoral bone marrow cells. FVB wild-type littermate mice were used as controls for FVB.A-ZIP/F1 fatless mice. During bone-marrow transplantation assays FVB or FVB.A-ZIP/F1 mice (CD45.1) received 200,000 MHC-matched (H^q) wild-type bone marrow cells from DBA/1J mice (CD45.2). Secondary transplants were performed through recovery of the CD45.2 DBA/2 bone marrow passed through the FVB wild-type or FVB.A-ZIP/F1 microenvironment, which was then

transplanted into FVB (CD45.1) recipients together with 250,000 FVB (CD45.1) wild-type competitor bone-marrow cells. For pharmacological inhibition of adipocyte formation, bone-marrow transplants were performed in wild-type female FVB mice as described above, except that 30 mg kg^{-1} BADGE or control vehicle (DMSO 10%) were administered in daily intraperitoneal injections starting 1 day before irradiation and continuing until day 14 after transplant. A total of 500 mg BADGE (Fluka) was re-suspended in 8.3 ml DMSO (Sigma) and diluted in PBS to a final concentration of 10% DMSO for administration at 30 mg kg^{-1} in 100 µl. Aliquots were stored at −20 °C and thawed daily. Multicolour flow cytometry was performed in a BD five-laser LSRII flow cytometer. Cell cycle analysis was performed with DAPI (Sigma) in cells fixed in 2% paraformaldehyde for 15 min at 4 °C. For all statistical analyses an unpaired two-tailed Student's *t*-test was performed assuming equal variance of samples. Error bars and confidence intervals represent the standard error of the mean (s.e.m.) unless otherwise indicated.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 6 January; accepted 27 April 2009.
Published online 10 June 2009.

1. Calvi, L. M. *et al.* Osteoblastic cells regulate the haematopoietic stem cell niche. *Nature* **425,** 841–846 (2003).
2. Zhang, J. *et al.* Identification of the haematopoietic stem cell niche and control of the niche size. *Nature* **425,** 836–841 (2003).
3. Kiel, M. J., Yilmaz, O. H., Iwashita, T., Terhorst, C. & Morrison, S. J. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121,** 1109–1121 (2005).
4. Bryon, P. A., Gentilhomme, O. & Fiere, D. Histomorphometric analysis of bone-marrow adipose density and heterogeneity in myeloid aplasia and dysplasia. *Pathol. Biol.* **27,** 209–213 (1979).
5. Moitra, J. *et al.* Life without white fat: a transgenic mouse. *Genes Dev.* **12,** 3168–3181 (1998).
6. Wright, H. M. *et al.* A synthetic antagonist for the peroxisome proliferator-activated receptor γ inhibits adipocyte differentiation. *J. Biol. Chem.* **275,** 1873–1877 (2000).
7. Neumann E.. Das Gesetz Verbreitung des gelben und rotten Markes in den Extremitätenknochen. *Zentbl. Med. Wiss.* **18,** 321–323 (1882).
8. Calvo, W., Fliedner, T. M., Herbst, E., Hügl, E. & Bruch, C. Regeneration of blood-forming organs after autologous leukocyte transfusion in lethally irradiated dogs. II. Distribution and cellularity of the marrow in irradiated and transfused animals. *Blood* **47,** 593–601 (1976).
9. Litten, M. & Orth, J. Ueber Veränderungen des Marks in Röhrenknochen unter verschiedenen pathologischen Verhältnissen. *Berl. Klin. Wschr.* **51,** 743–751 (1877).
10. Christensen, J. L. & Weissman, I. L. Flk-2 is a marker in hematopoietic stem cell differentiation: a simple method to isolate long-term stem cells. *Proc. Natl Acad. Sci. USA* **98,** 14541–14546 (2001).
11. Osawa, M., Hanada, K., Hamada, H. & Nakauchi, H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* **273,** 242–245 (1996).
12. Pan, Z. *et al.* Effects of hindlimb unloading on *ex vivo* growth and osteogenic/adipogenic potentials of bone marrow-derived mesenchymal stem cells in rats. *Stem Cells Dev.* **17,** 795–804 (2008).
13. Yang, L. *et al.* Identification of $Lin^{-}Sca1^{+}kit^{+}CD34^{+}Flt3^{-}$ short-term hematopoietic stem cells capable of rapidly reconstituting and rescuing myeloablated transplant recipients. *Blood* **105,** 2717–2723 (2005).
14. Tavassoli, M., Maniatis, A. & Crosby, W. H. Induction of sustained hemopoiesis in fatty marrow. *Blood* **43,** 33–38 (1974).
15. Botolin, S. & McCabe, L. R. Inhibition of PPARγ prevents type I diabetic bone marrow adiposity but not bone loss. *J. Cell. Physiol.* **209,** 967–976 (2006).
16. Furuhashi, M. *et al.* Treatment of diabetes and atherosclerosis by inhibiting fatty-acid-binding protein aP2. *Nature* **447,** 959–965 (2007).
17. Digman, C., Klein, A. K. & Pittas, A. G. Leukopenia and thrombocytopenia caused by thiazolidinediones. *Ann. Intern. Med.* **143,** 465–466 (2005).
18. Maaravi, Y. & Stessman, J. Mild, reversible pancytopenia induced by rosiglitazone. *Diabetes Care* **28,** 1536 (2005).
19. Berria, R. *et al.* Reduction in hematocrit and hemoglobin following pioglitazone treatment is not hemodilutional in Type II diabetes mellitus. *Clin. Pharmacol. Ther.* **82,** 275–281 (2007).
20. Lazarenko, O. P. *et al.* Rosiglitazone induces decreases in bone mass and strength that are reminiscent of aged bone. *Endocrinology* **148,** 2669–2680 (2007).
21. Nishikawa, M. *et al.* Changes in hematopoiesis-supporting ability of C3H10T1/2 mouse embryo fibroblasts during differentiation. *Blood* **81,** 1184–1192 (1993).
22. Corre, J. *et al.* Human subcutaneous adipose cells support complete differentiation but not self-renewal of hematopoietic progenitors. *J. Cell. Physiol.* **208,** 282–288 (2006).
23. Belaid-Choucair, Z. *et al.* Human bone marrow adipocytes block granulopoiesis through neuropilin-1-induced granulocyte colony-stimulating factor inhibition. *Stem Cells* **26,** 1556–1564 (2008).
24. Miharada, K. *et al.* Lipocalin 2-mediated growth suppression is evident in human erythroid and monocyte/macrophage lineage cells. *J. Cell. Physiol.* **215,** 526–537 (2008).
25. Yan, Q. W. *et al.* The adipokine lipocalin 2 is regulated by obesity and promotes insulin resistance. *Diabetes* **56,** 2533–2540 (2007).
26. Yokota, T. *et al.* Adiponectin, a new member of the family of soluble defense collagens, negatively regulates the growth of myelomonocytic progenitors and the functions of macrophages. *Blood* **96,** 1723–1732 (2000).
27. Zhang, Y. *et al.* Tumor necrosis factor (TNF) is a physiologic regulator of hematopoietic progenitor cells: increase of early hematopoietic progenitor cells in TNF receptor p55-deficient mice *in vivo* and potent inhibition of progenitor cell proliferation by TNFα *in vitro*. *Blood* **86,** 2930–2937 (1995).
28. Hotamisligil, G. S., Shargill, N. S. & Spiegelman, B. M. Adipose expression of tumor necrosis factor-α: direct role in obesity-linked insulin resistance. *Science* **259,** 87–91 (1993).
29. DiMascio, L. *et al.* Identification of adiponectin as a novel hemopoietic stem cell growth factor. *J. Immunol.* **178,** 3511–3520 (2007).
30. Nuttall, M. E. & Gimble, J. M. Controlling the balance between osteoblastogenesis and adipogenesis and the consequent therapeutic implications. *Curr. Opin. Pharmacol.* **4,** 290–294 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Lazo-Kallanian, J. Daley, G. Losyev and R. Mathieu for assistance with flow cytometry; R. Bronson for assistance with pathological analysis; P. Dunning, E. Snay and S. Carlton for assistance with small animal imaging; S. Loewer for translation of historical references; and S. McKinney-Freeman, A. Yabuuchi, K. Ng and R. Chapman for mouse and technical assistance. O.N. was partially funded by the Barrie de la Maza Foundation. P.L.W. was supported by a Hematology Training Grant from the National Institutes of Health (NIH T32- HL -7623). G.Q.D. was supported by grants from the NIH and the NIH Director's Pioneer Award of the NIH Roadmap for Medical Research. G.Q.D. is the recipient of the Clinical Scientist Award in Translational Research from the Burroughs Wellcome Fund and the Leukemia and Lymphoma Society, and is an Investigator of the Howard Hughes Medical Institute.

Author Contributions O.N and G.Q.D. conceived the original idea, designed experiments and wrote the manuscript. O.N., V.N. and P.L.W. performed experiments and analysed results. P.V.H. contributed to stromal differentiation essays. O.N. and F.F. performed quantitative acquisition and analysis of mCT and mPET. All authors edited and reviewed the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.Q.D. (george.daley@childrens.harvard.edu).

METHODS

Animals. All mice were purchased from Jackson Laboratories and sex-, weight- and age-matched or breeding colonies were established in house. Experiments were carried out with IACUC approval from CHB.

Bone-marrow preparations. Femurs, thorax and tails were isolated free of muscle and tendons; when appropriate, the spinal cord was carefully removed. Bones were crushed in IMDM with mortar and pestle, filtered through a 70-μm filter and washed with PBS. A sample was removed, stained with CD45–FITC (1:200) and 7-amino-actinomycin D (1:100) in 50 μl and the volume raised to 500 μl with PBS and reference beads (Sigma). A viable $CD45^+$ cell count was then obtained with a F500-Coulter flow cytometer. Red blood cells from bone marrow in homeostatic conditions (pre-transplant) were lysed with RBC lysis buffer (Sigma). For early post-transplant analysis, bone marrow was not lysed and efforts were made to perform minimal manipulation before c.f.u. plating or secondary transplantation (red blood cells were excluded as counts were based on $CD45^+$ cells only).

Flow cytometry. Multicolour analysis for progenitor and stem-cell quantification was performed on a FACS-Aria+UV or on a 5-laser-LSRII flow cytometer (Beckton-Dickinson). Cells were stained in PBS 2% FCS for 1 h with CD34–FITC (1:50, Beckton-Dickinson), Flk2–PE (1:100, Beckton-Dickinson), lineage cocktail-PECy5 (Ter119/B220/CD19/CD3/CD4/CD8/Nk1.1 from eBiosciences; mixed 1:1 except CD3 (2:1), and used 1:200), FcγRIII/II-PECy7 (1:200, Biolegend), cKit-APC (1:200, Beckton-Dickinson), CD45.1- or CD45.2-APCCy7 (1:100, Biolegend), CD45-biotin (1:200, Beckton-Dickinson), streptavidin-Pacific orange (1:2000, Invitrogen), and Sca1-Pacific blue (1:100 Biolegend). For cell cycle analysis, bone-marrow cells were stained in cold 2% IFS and fixed in cold 2% PFA for 15 min, then washed, stained in DAPI solution for 10 min at room temperature (0.1% (v/v) Triton X-100 and 1 μg ml^{-1} DAPI in PBS[31,32]), washed and immediately analysed. Ki-67 analysis was done as described previously[33]. Bone-marrow cells were stained by cell surface markers (Sca1-FITC (1:100, Beckton-Dickinson), lineage-cocktail (1:200), FcγRIII/II-PECy7 (1:200, Abcam), cKit-APC (1:200), CD45-APCCy7 (1:100, BD), CD150-biotin (1:300, BioLegend), streptavidin-Alexa Fluor 680 (1:200, Invitrogen)), washed, fixed in cold Cytofix/Cytoperm buffer (Beckton-Dickinson) for 20 min, then washed and incubated in Permwash buffer (Beckton-Dickinson) with Ki-67-PE (1:100, Beckton-Dickinson) for 5 h at 4 °C. Cells were treated with 1 μg ml^{-1} DAPI for 10 min, washed and resuspended in PBS 2% FCS immediately before acquisition.

Progenitor assays. Colony-forming unit (c.f.u.) assays were performed in complete M3434 methylcellulose (Stem Cell Technologies) following the manufacturer's instructions. Colonies were scored on day 8–10 on coded plates for unbiased counts.

Bone-marrow transplantation. Mice were lethally irradiated with 11–12 Gy split-dose 2.5 h apart, and bone-marrow transplants performed within 24 h by tail-vein injection. Engraftment was measured monthly through eye bleed and FACS analysis with CD45.1-FITC (eBiosciences), CD45.2-PE or CD45.2-biotin, CD3-PE, CD19-PE, Mac-1-PE, Gr1-PE, F4/80-APC. All antibodies were ordered from Beckton-Dickinson unless otherwise specified. Mice for which engraftment was below 0.5% were considered non-engrafted and were not taken into account for calculation of competitive repopulation units.

Stromal cell culture and differentiation. OP9 cells (ATCC) were expanded in MEMα 15% IFS + penicillin/streptomycin/glutamine (Gibco); media changed every 3–4 days. Adipocytic differentiation was performed on confluent OP9 cells plated at 10,000 cm^2 overnight, then treated with 0.5 mM isobutylmethylxantine (IBMX, Sigma; 1,000× stock in DMSO), 5μg ml^{-1} insulin (Sigma; 1,000× stock in PBS) and 10^{-6} M dexamethasone (Sigma; 1,000× stock in ethanol) on the first week, then maintained with insulin and dexamethasone only for another 10 days. Media changes were made every 3–4 days with fresh aliquots maintained in the dark at −20 °C. Haematopoietic co-cultures with FACS-sorted $ckit^+Lin^-Sca1^+$ (KLS) haematopoietic stem cells were performed in IMDM 10% IFS at 37 °C and 5% CO_2; media was changed on the seventh day of co-culture. For co-cultures, 2,000 KLS cells were plated per well. Transwell assays used 12-mm polyester inserts with 0.4-μm pore size (Corning) with adipocytic or undifferentiated OP9 cells plated in the bottom of the plate while undifferentiated OP9 cells were co-cultured with KLS in the upper insert.

mPET analysis. For micro-positron emission tomography (mPET), mice were administered equal doses (5.6–9.3 MBq) of sodium fluoride-18 (^{18}F) by tail-vein injection. Animals were imaged before bone-marrow transplantation, and the same cohort of animals was re-assayed at different times after transplant. Exactly 30 min after the ^{18}F injection, mice were imaged with a Focus 120 microPET scanner (Siemens). For normalization, the total dose injected was determined immediately before and after imaging by introducing the anaesthetized mouse in the dose calibrator and calculating the mid-acquisition dose as the average of the pre- and post-acquisition measurements adjusted by time decay. ASIPro software (Siemens) was then used for image analysis. To determine the ^{18}F uptake in tibias and tails, a three-dimensional ROI (region of interest) was created in transverse sections through the selection of six 9 × 9 pixel planes moving distally from the tibial plate or 12 planes moving distally from the beginning of the free tail. The mean dose in the ROI (Bq ml^{-1}) was then normalized to the total injected dose and the ratios from pre- to post-transplant ^{18}F uptake were then calculated. When the whole cohort of mice could not be assayed on the same day after transplant, data from the two closest dates were interpolated.

mCT analysis. For high-resolution micro-computerized tomography (mCT) analysis, tibias were dissected, fixed in 10% formalin and imaged with the Siemens microCAT II system using a 22.75 μm pixel size. Beam angle of increment was 1 degree and tube voltage and current were 80 kVp and 450 μA per view as described previously[15]. Each run included wild-type and fatless bones and a calibration phantom to calibrate greyscale values with respect to mass density and maintain consistency. A fixed density threshold was established through normalization to the calibration density phantom to separate trabecular bone from bone marrow (1,350 mg ml^{-1}). Trabecular bone analyses were done with ImageJ and AMIRA image analysis software in a 0.2-mm-thick three-dimensional square region of trabecular bone defined at 1 mm under the growth plate of the proximal tibia extending 0.9 mm towards the diaphysis and excluding the outer cortical shell.

31. Pozarowski, P. & Darzynkiewicz, Z. Analysis of cell cycle by flow cytometry. *Methods Mol. Biol.* **281**, 301–311 (2004).
32. Goodell, M. A., Brose, K., Paradis, G., Conner, A. S. & Mulligan, R. C. Isolation and functional properties of murine hematopoietic stem cells that are replicating *in vivo*. *J. Exp. Med.* **183**, 1797–1806 (1996).
33. Wilson, A. *et al.* c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes Dev.* **18**, 2747–2763 (2004).

LETTERS

CD14 regulates the dendritic cell life cycle after LPS exposure through NFAT activation

Ivan Zanoni[1], Renato Ostuni[1], Giusy Capuano[1], Maddalena Collini[2], Michele Caccia[2], Antonella Ellena Ronchi[1], Marcella Rocchetti[1], Francesca Mingozzi[1], Maria Foti[1], Giuseppe Chirico[2], Barbara Costa[1], Antonio Zaza[1], Paola Ricciardi-Castagnoli[1]† & Francesca Granucci[1]

Toll-like receptors (TLRs) are the best characterized pattern recognition receptors[1]. Individual TLRs recruit diverse combinations of adaptor proteins, triggering signal transduction pathways and leading to the activation of various transcription factors, including nuclear factor κB, activation protein 1 and interferon regulatory factors[2]. Interleukin-2 is one of the molecules produced by mouse dendritic cells after stimulation by different pattern recognition receptor agonists[3–6]. By analogy with the events after T-cell receptor engagement leading to interleukin-2 production, it is therefore plausible that the stimulation of TLRs on dendritic cells may lead to activation of the Ca^{2+}/calcineurin and NFAT (nuclear factor of activated T cells) pathway. Here we show that mouse dendritic cell stimulation with lipopolysaccharide (LPS) induces Src-family kinase and phospholipase Cγ2 activation, influx of extracellular Ca^{2+} and calcineurin-dependent nuclear NFAT translocation. The initiation of this pathway is independent of TLR4 engagement, and dependent exclusively on CD14. We also show that LPS-induced NFAT activation via CD14 is necessary to cause the apoptotic death of terminally differentiated dendritic cells, an event that is essential for maintaining self-tolerance and preventing autoimmunity[7,8]. Consequently, blocking this pathway *in vivo* causes prolonged dendritic cell survival and an increase in T-cell priming capability. Our findings reveal novel aspects of molecular signalling triggered by LPS in dendritic cells, and identify a new role for CD14: the regulation of the dendritic cell life cycle through NFAT activation. Given the involvement of CD14 in disease, including sepsis and chronic heart failure[9,10], the discovery of signal transduction pathways activated exclusively via CD14 is an important step towards the development of potential treatments involving interference with CD14 functions.

Wild-type LPS, termed smooth LPS (the form of LPS used here), comprises three covalently linked moieties: lipid A, rough core oligosaccharide and an O-antigenic side chain determining serotype specificity. The TLR4–MD2 complex and the CD14 co-receptor are necessary and sufficient for a full response to smooth LPS[11]. CD14 is expressed by both $CD8\alpha^+$ and $CD8\alpha^-$ mouse dendritic cells, although at lower levels than expressed by macrophages (Supplementary Fig. 1). We tested whether LPS-induced interleukin-2 (IL-2) production by dendritic cells was dependent on NFAT. Ca^{2+} mobilization is the first event in NFAT activation. We observed that LPS was able to induce Ca^{2+} flux in bone-marrow-derived dendritic cells (BMDCs) and in a homogeneous dendritic cell line, D1 cells[12] (Fig. 1a, Supplementary Fig. 2 and Supplementary Movies). Next, we investigated the features of Ca^{2+} mobilization by LPS in BMDCs in detail, by recording Ca^{2+} transients in individual cells. The intracellular calcium concentration ($[Ca^{2+}]_i$) was increased by LPS and ATP ($P < 0.05$ versus untreated cells, Fig. 1b). Experiments in Ca^{2+}-free medium demonstrated that the increase in $[Ca^{2+}]_i$ induced by LPS was due to the influx of extracellular Ca^{2+} (Fig. 1c, responding cells in Ca^{2+}-free medium less than 1%). To determine the molecular component of the LPS receptor responsible for Ca^{2+} flux directly, we measured Ca^{2+} influx in TLR4- and CD14-deficient BMDCs. Ca^{2+} transients were fully preserved in TLR4-deficient BMDCs ($43 \pm 6\%$ of cells responding, Fig. 1d), whereas they were completely abolished in CD14-deficient BMDCs ($\leq 1\%$ of cells responding, Fig. 1d). To confirm that TLR4 is not involved in this process, Ca^{2+} mobilization was tested in each of the following mutant cell types: MyD88-, MD2 (also called LY96)- and TRIF (also called TICAM1)-deficient, and MyD88/TRIF- and TLR4/TLR2-double-deficient BMDCs. In all of these mutant cells Ca^{2+} mobilization was fully preserved (Supplementary Fig. 3), confirming that CD14 was the receptor responsible for Ca^{2+} influxes after LPS activation. Also, there was no Ca^{2+} mobilization in BMDCs in response to taxol (responding cells $\leq 1\%$), a selective TLR4 agonist[13] (Fig. 1d); this is further confirmation of the involvement of CD14. We next conducted various tests to ensure that the results obtained concerning Ca^{2+} transients could not have been due to the presence of contaminants in our LPS preparations (see Supplementary Fig. 4).

CD14 is a glycosylphosphatidylinositol-anchored receptor (GPI-AR). Cross-linking of GPI-ARs, such as CD59 (ref. 14), triggers the activation of the intracellular inositol-1,4,5-trisphosphate ($Ins(1,4,5)P_3$) Ca^{2+} pathway through the association with lipid rafts and Src kinase activation[15]. Therefore, we investigated whether CD14 membrane localization, presence of lipid rafts and Src kinase activation were conditions required to obtain Ca^{2+} mobilization. As shown in Fig. 1e, soluble CD14 did not restore extracellular Ca^{2+} influx in CD14-deficient dendritic cells, and disruption of lipid rafts by cholesterol depletion totally impaired Ca^{2+} mobilization. Thus, consistent with previous results[16], CD14 localization in lipid rafts[17] seems to be required for LPS-induced Ca^{2+} mobilization. Moreover, Src kinase inhibition completely abolished Ca^{2+} mobilization (Fig. 1e), indicating direct involvement of Src kinases in this process.

The induction of Ca^{2+} transients by LPS was also fully dependent on phospholipase C (PLC)-γ2 (responding PLC-γ2-deficient BMDCs $\leq 1\%$; Fig. 1e). Consistent with these results, PLC-γ2 activation was observed in wild-type but not in CD14-deficient BMDCs on LPS exposure (Supplementary Fig. 5).

We then analysed NFAT activation after LPS treatment in wild-type and CD14-deficient BMDCs. As shown in Fig. 2a, NFATc2

[1]Department of Biotechnology and Bioscience, University of Milano-Bicocca, Piazza della Scienza 2, 20126 Milan, Italy. [2]Department of Physics, University of Milano-Bicocca, Piazza della Scienza 3, 20126 Milan, Italy. †Present address: Singapore Immunology Network (SIgN), Biomedical Sciences Institutes, Agency for Science, Technology and Research (A*STAR), 8A Biomedical Grove, IMMUNOS, 138648, Singapore.

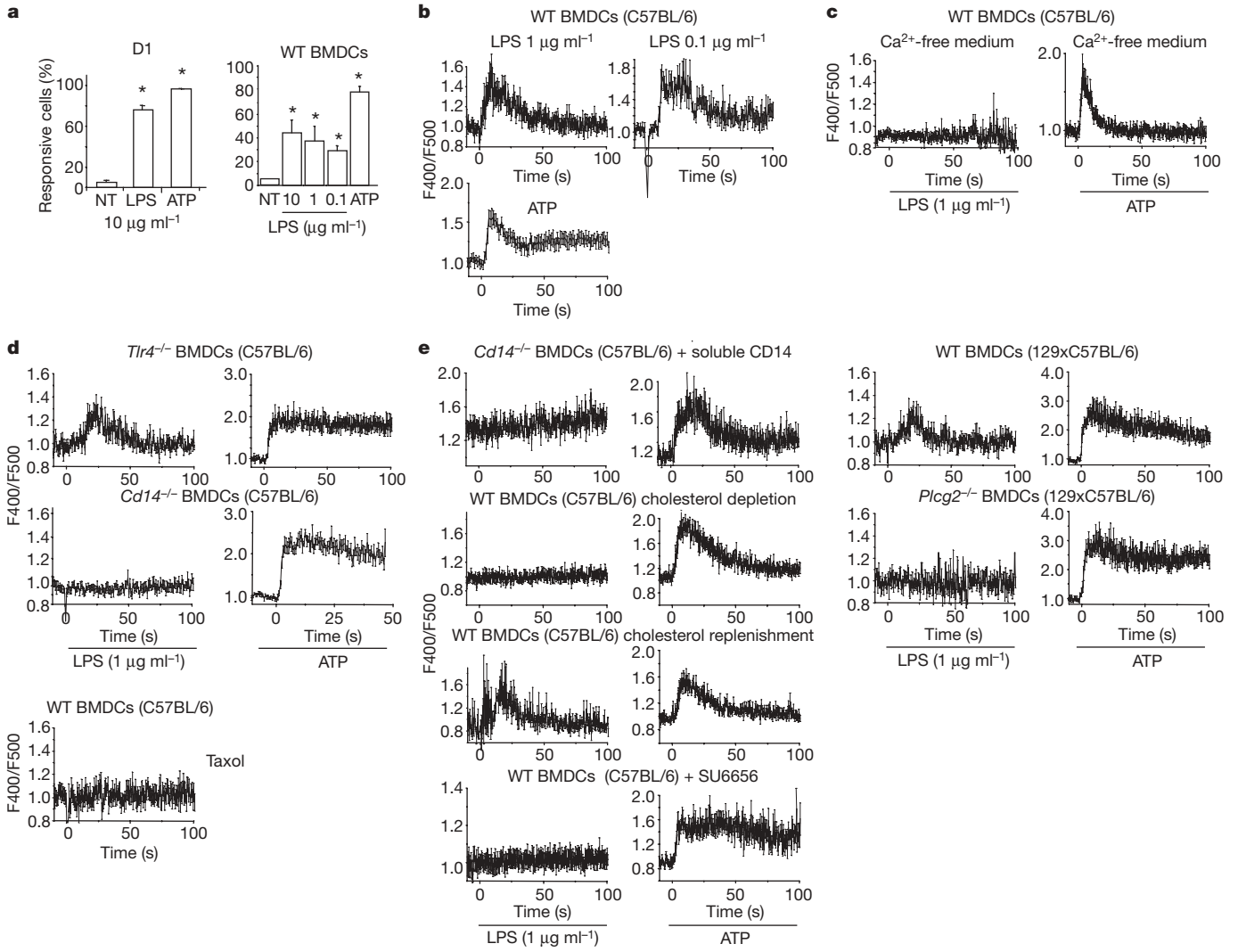


Figure 1 | CD14-dependent Ca^{2+} mobilization in dendritic cells following LPS treatment. **a**, Percentage of D1 cells and BMDCs showing Ca^{2+} mobilization in the absence (NT) and presence of LPS or ATP. Asterisk, $P < 0.05$ versus NT. **b–e**, Ca^{2+} transients in wild-type and mutant BMDCs and wild-type BMDCs in the presence of the Src kinase inhibitor (SU6656, 260 nM) or soluble CD14 (2 μg ml^{-1}). In **e** Ca^{2+} transients were recorded after cholesterol depletion and replenishment. Point 0 indicates the time of stimulus administration. Means and s.d. for a minimum of 30 cells are shown. Experiments were repeated at least three times. The ratio of fluorescence emissions at 400 nm/40 nm band-pass to those at 500 nm/20 nm band-pass was recorded (F400/F500) and used as an index of $[Ca^{2+}]_i$.

nuclear translocation occurred in wild-type but not in CD14-deficient BMDCs. In agreement with the Ca^{2+} mobilization studies, NFAT was translocated to the nucleus in wild-type, TLR4-, MD2- and MyD88/TRIF-deficient BMDCs but not in CD14- and PLC-γ2-deficient BMDCs or in wild-type BMDCs stimulated with LPS in the presence of the Src kinase inhibitor (Supplementary Fig. 6). These observations demonstrate that CD14 is directly responsible for activating, in dendritic cells, a TLR4-independent, Src-kinase- and PLC-γ2-dependent transduction pathway leading to NFAT activation. This is in addition to CD14 enhancing cellular responses to low doses of LPS[18,19] and being required for the LPS-induced recruitment of TRAM and TRIF[20].

We observed that, diversely from macrophages[21] (Fig. 2b), the production of tumour-necrosis factor (TNF)-α and IL-6 by dendritic cells after LPS stimulation was TRIF-independent (Fig. 2b) and, consequently, it was also CD14-independent[11,20], at least at high LPS doses (Fig. 2b). It is therefore possible that the selective engagement of TLR4 without CD14 activation allows the production of NF-κB-dependent and NFAT-independent cytokines, such as TNF-α[22] and IL-6[23], whereas it selectively prevents the production of cytokines, such as IL-2, that are transcriptionally dependent on NFAT, in addition to NF-κB and activation protein 1 (AP1).

We analysed the behaviour of CD14-deficient BMDCs after LPS stimulation and of wild-type BMDCs after activation with the TLR4-selective stimulus taxol to confirm the existence of a CD14-specific pathway. IL-2 production was impaired after the stimulation of CD14-deficient BMDCs with various concentrations of LPS, including very high concentrations at which the production of TNF-α and IL-6 was largely preserved (Fig. 2c). Similarly, the stimulation of wild-type BMDCs with taxol induced the production of TNF-α and IL-6, but not IL-2 (Fig. 2c). Nevertheless, coupling TLR4 stimulation with Ca^{2+} mobilization, using thapsigargin, restored IL-2 production (Fig. 2d and Supplementary Fig. 8). Similar results were obtained *in vivo* (Supplementary Fig. 8).

We used various inhibitors blocking the NFAT pathway to confirm CD14-dependent activation of the Ca^{2+}/calcineurin pathway induced by LPS. In particular, inhibition of the influx of external Ca^{2+} by the Ca^{2+} chelator EGTA selectively impaired IL-2 production but did not downregulate TNF-α or IL-6 synthesis (Supplementary Fig. 9a). This was also observed in the presence of thapsigargin (Supplementary Fig. 9b). Moreover, inhibition of Src kinases, PLC-γ2 and calcineurin downregulated IL-2 production, although TNF-α and IL-6 production was unaffected (Supplementary Fig. 9a, c). We also used the VIVIT peptide[24] to

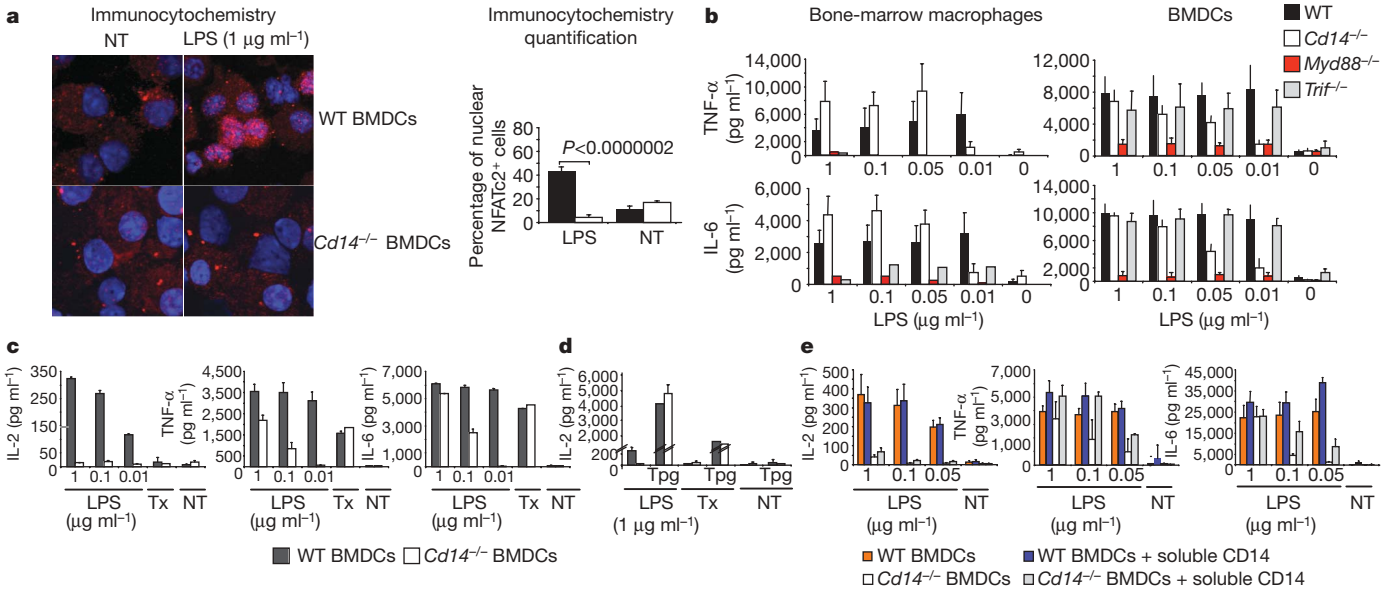


Figure 2 | CD14-dependent activation of NFAT in BMDCs after LPS treatment. **a**, NFATc2 translocation before (NT) and after LPS challenge for 2 h. Blue, DAPI; red, anti-NFATc2; original magnification, ×630. Percentages represent means and standard error of five fields (~50 cells each) in two experiments. **b**, LPS-induced TNF-α and IL-6 production by bone-marrow macrophages and BMDCs. **c**, LPS- or taxol (Tx, 100 μM)-induced cytokine production by BMDCs. **d**, LPS- or taxol-induced IL-2 production by BMDCs with and without thapsigargin (Tpg, 50 nM). **e**, LPS-induced cytokine production by BMDCs with and without soluble CD14 (2 μg ml^{-1}). **b**, **c**, **d**, **e** are representative of three experiments. Error bars in **b**–**e** represent standard errors.

inhibit NFAT proteins and obtained similar results (Supplementary Fig. 10).

In agreement with the Ca^{2+} mobilization data, addition of soluble CD14 to CD14-deficient BMDCs restored TNF-α production, as already observed in macrophages[20], but not IL-2 production at either low or high LPS concentrations (Fig. 2e). This observation supports the notion that a membrane localization of CD14 is required to activate the Ca^{2+} pathway.

As NFAT proteins are major regulators of many biological processes, we investigated the physiological role of CD14-mediated NFAT activation in dendritic cells, in addition to IL-2 production. We performed a kinetic microarray analysis to identify genes modulated specifically by NFAT in LPS-treated dendritic cells (see Supplementary Fig. 11 for details). From this analysis we hypothesized that LPS-activated dendritic cells in conditions allowing NFAT activation were exhausted cells with little or falling gene transcription and protein translation; in contrast, dendritic cells activated with LPS in conditions inhibiting NFAT activation were transcriptionally and translationally active, presumably not exhausted and not undergoing apoptosis (see Supplementary Table 1 for details).

To test whether CD14-mediated activation of the NFAT pathway was required to induce apoptotic death of terminally differentiated

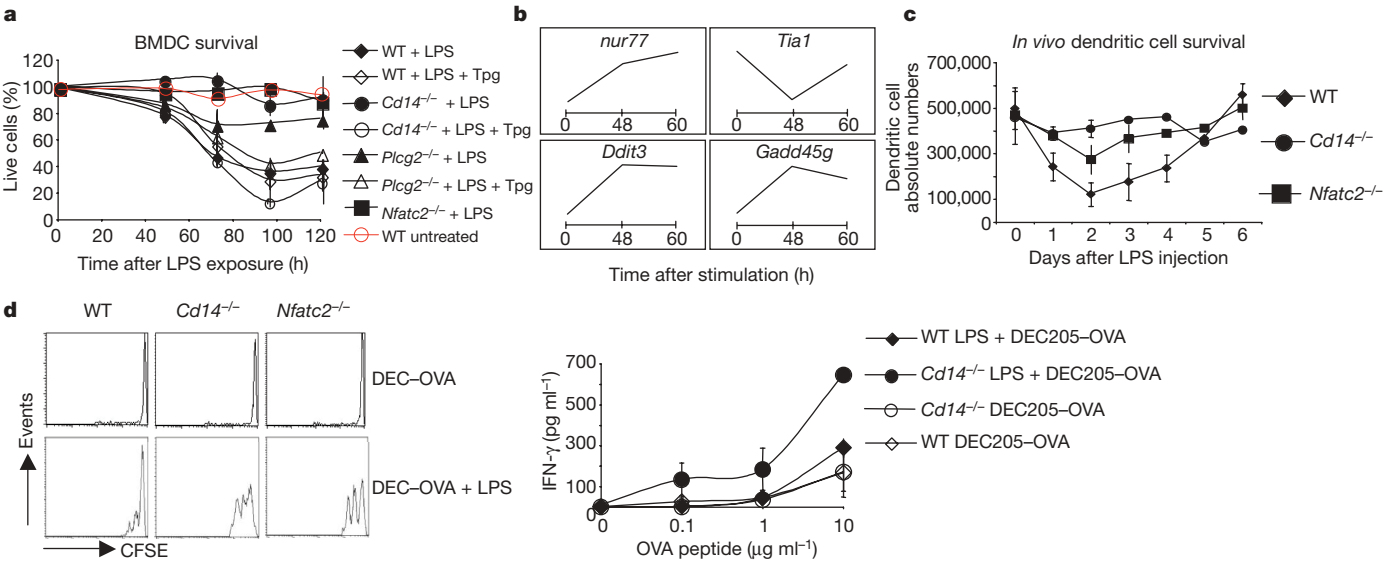


Figure 3 | Ca^{2+}-NFAT-mediated regulation of dendritic cell death. **a**, BMDC survival after incubation with LPS (1 μg ml^{-1}, $P < 0.00001$, wild-type versus mutant BMDCs) or LPS plus thapsigargin (50 nM). Survival of unstimulated cells kept in culture with granulocyte–macrophage colony-stimulating factor (GM-CSF) is also shown. **b**, Apoptosis genes regulated by NFAT after LPS stimulation. **c**, Absolute numbers of CD11c^{+} cells in spleens after intravenous LPS (1 μg g^{-1}) injection ($P \leq 0.0006$, wild-type versus mutant mice). Data are representative of two independent experiments (four mice per group). **d**, Left panel: antigen-specific T-cell proliferation *in vivo*. Right panel: IFN-γ production after *in vitro* re-stimulation. The experiment was repeated twice with similar results. Error bars in **a**, **c**, **d** represent standard errors.

cells, wild-type BMDCs and CD14-, NFATc2- and PLC-γ2-deficient BMDCs were activated with LPS and the percentage of living cells measured at various times thereafter. The c2 isoform of NFAT was chosen because it has a pro-apoptotic function in T and B cells[25]. Less than 30% of wild-type BMDCs survived after activation, whereas 70–100% of CD14-, NFATc2- and PLC-γ2-deficient BMDCs were still alive five days after exposure to LPS (Fig. 3a). CD14- and PLC-γ2-deficient BMDC death after LPS activation was restored by thapsigargin (Fig. 3a); thapsigargin alone had no effect (data not shown). Similar results were obtained using different LPS doses and LPS from different sources (Supplementary Fig. 12). A new microarray experiment performed at later time points (48 and 60 h) allowed us to identify four pro-apoptotic genes specifically modulated by NFAT: *nur77*, *Gadd45g*, *Ddit3* (also called *gadd153* and *CHOP-10*) and *Tia1* (Fig. 3b, see Supplementary Table 2 for details).

In vivo, splenic dendritic cell numbers declined quickly in wild-type C57BL/6 mice upon LPS administration, whereas absolute dendritic cell numbers remained almost constant in CD14-deficient and NFATc2-deficient mice after the same treatment (Fig. 3c). Although we cannot exclude an indirect effect *in vivo*, these findings provide further evidence that NFAT pathway activation via CD14 is required for dendritic cell apoptosis. The dendritic cell decline observed in NFATc2-deficient mice may have been the consequence of the effect of other NFAT isoforms with overlapping functions.

Apoptotic death is the terminal step of fully matured dendritic cells. This process is essential for regulating peripheral tolerance because dendritic cell accumulation resulting from apoptosis deficiency can induce over-activation of responder lymphocytes, resulting in systemic autoimmunity[7,8]. We tested whether blockade of the CD14/NFAT pathway leading to dendritic cell persistence after activation could be responsible for an increase in T-cell priming *in vivo*. Wild-type, CD14- and NFATc2-deficient mice were injected with LPS and anti-DEC205–OVA conjugate[26] and 8 h later received carboxyfluorescein diacetate succinimidyl ester (CFSE)-labelled $CD4^+$ T cells from OVA-specific OT-II mice. As shown in Fig. 3d, T-cell proliferation *in vivo* was stronger in mutant than wild-type mice and, consistently, the capacity to produce IFN-γ after re-stimulation *in vitro* was stronger when T cells were recovered from mutant than wild-type mice. Thus, mutant mice had an enhanced capacity to induce T-cell-specific immune responses.

Diversely from dendritic cells, macrophages do not die after activation with LPS but simply become refractory to a further re-challenge[27]. Tissue-resident macrophage survival after activation is crucial for inflammation resolution[28]. We asked whether macrophage survival after LPS encounter was due to their inability to activate the Ca^{2+} pathway. Bone-marrow-derived macrophages were unable to mobilize Ca^{2+} (Fig. 4a). Moreover, Supplementary Fig. 13 shows that after LPS encounter NFAT activation and NFAT-dependent gene transcription were not measurable in macrophages. To investigate whether the lack of Ca^{2+}-NFAT pathway activation in macrophages was responsible for their survival upon LPS encounter, we induced Ca^{2+} mobilization with thapsigargin concomitantly with LPS stimulation. In these conditions, bone-marrow macrophage death could be induced in an NFATc2-dependent manner (Fig. 4b). In addition, *in vivo* splenic macrophages (identified as $CD11b^+CD11c^{dim}$ cells) did not decline in number after intravenous injection of LPS (Fig. 4b).

Using real-time quantitative PCR we analysed the expression modulation in bone-marrow macrophages of *nur77*, one of the best-characterized pro-apoptotic genes selected by microarray analysis in dendritic cells. As shown in Fig. 4c, *nur77* was not modulated in macrophages after LPS stimulation. Nevertheless, *nur77* messenger RNA upregulation could be induced in an NFAT-dependent manner after macrophage co-stimulation with LPS and thapsigargin. Moreover, in this validation experiment, we also confirmed that the upregulation of *nur77* mRNA in dendritic cells after LPS stimulation was dependent on NFATc2 (Fig. 4c).

These observations further corroborate our findings concerning the functional consequences of CD14-mediated Ca^{2+} influx and NFAT activation in dendritic cells (Supplementary Fig. 14). Nevertheless, they highlight significant differences in the signal transduction pathways induced by LPS in dendritic cells and macrophages.

By revealing novel aspects of molecular signalling triggered by LPS in dendritic cells, this work provides insights into the complexity of cellular responses to bacterial infections.

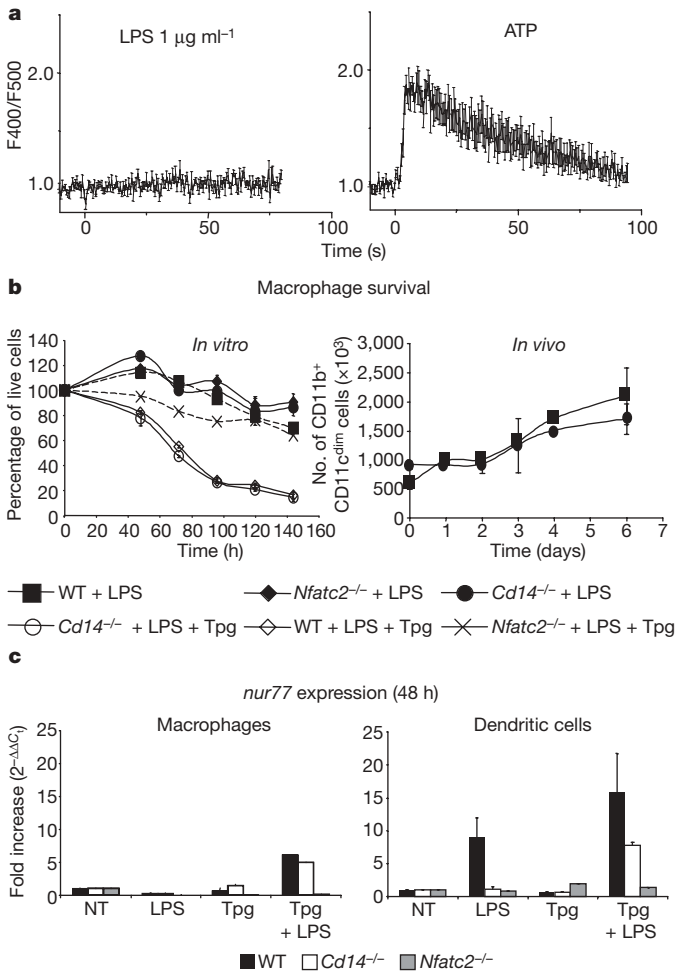


Figure 4 | Ca^{2+} mobilization and survival of macrophages after LPS treatment. **a**, Ca^{2+} transients in bone marrow macrophages. Point 0 indicates time of stimulus administration. Means and s.d. of 30 cells minimum are shown. Experiments were repeated at least three times. **b**, Left panel: survival of wild-type and mutant bone-marrow macrophages after incubation with LPS (1 μg ml^{-1}) or LPS plus thapsigargin (50 nM). Right panel: absolute numbers of $CD11b^+CD11c^{dim}$ cells in spleens of wild-type and CD14-deficient mice after intravenous LPS injection (30 μg). Data represent two independent experiments (four mice per group). **c**, Fold increase of *nur77* expression in stimulated versus unstimulated wild-type and mutant cells. Error bars in **b** and **c** represent standard errors.

METHODS SUMMARY

Calcium measurements. $[Ca^{2+}]_i$ was determined by a fluorometric ratio technique. Cells were loaded with 2 μM indo-1 (Molecular Probes) by incubation at 37 °C for 20 min. Cells were then washed three times with PBS to allow for intracellular de-esterification of indo-1. A direct optical microscope (Olympus, BX51) with a two-photon Ti-Sapphire laser source (720-nm wavelength; Mai Tai, SpectraPhysics) was used for indo-1 excitation. The fluorescence signals emitted by indo-1-loaded cells were digitized at 200 Hz and recorded every 0.5–0.8 s. The ratio of fluorescence emissions at 400 nm/40 nm band-pass to those at 500 nm/20 nm band-pass was recorded (R400/500) and used as an index of $[Ca^{2+}]_i$. Data were normalized to baseline. This approach overcame

possible problems of uncertainty related to the calibration of fluorescent Ca^{2+} indicators.

In some cases, cells were analysed in calcium-free PBS or calcium-free PBS supplemented with thapsigargin (50 nM).

Images of Ca^{2+} responses were measured by a laser-scanning confocal system using the membrane-permeable dye Fluo4-AM.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 March; accepted 28 April 2009.
Published online 14 June 2009.

1. Akira, S. Toll-like receptors and innate immunity. *Adv. Immunol.* **78,** 1–56 (2001).
2. Kaisho, T. & Akira, S. Toll-like receptor function and signaling. *J. Allergy Clin. Immunol.* **117,** 979–987 (2006).
3. Granucci, F., Feau, S., Angeli, V., Trottein, F. & Ricciardi-Castagnoli, P. Early IL-2 production by mouse dendritic cells is the result of microbial-induced priming. *J. Immunol.* **170,** 5075–5081 (2003).
4. Guiducci, C., Valzasina, B., Dislich, H. & Colombo, M. P. CD40/CD40L interaction regulates $CD4^+CD25^+$ T_{reg} homeostasis through dendritic cell-produced IL-2. *Eur. J. Immunol.* **35,** 557–567 (2005).
5. Rogers, N. C. *et al.* Syk-dependent cytokine induction by Dectin-1 reveals a novel pattern recognition pathway for C type lectins. *Immunity* **22,** 507–517 (2005).
6. Yamazaki, S. *et al.* Direct expansion of functional $CD25^+$ $CD4^+$ regulatory T cells by antigen-processing dendritic cells. *J. Exp. Med.* **198,** 235–247 (2003).
7. Chen, M. *et al.* Dendritic cell apoptosis in the maintenance of immune tolerance. *Science* **311,** 1160–1164 (2006).
8. Stranges, P. B. *et al.* Elimination of antigen-presenting cells and autoreactive T cells by Fas contributes to prevention of autoimmunity. *Immunity* **26,** 629–641 (2007).
9. Van Amersfoort, E. S., Van Berkel, T. J. & Kuiper, J. Receptors, mediators, and mechanisms involved in bacterial sepsis and septic shock. *Clin. Microbiol. Rev.* **16,** 379–414 (2003).
10. Genth-Zotz, S. *et al.* The anti-CD14 antibody IC14 suppresses *ex vivo* endotoxin stimulated tumor necrosis factor-α in patients with chronic heart failure. *Eur. J. Heart Fail.* **8,** 366–372 (2006).
11. Beutler, B. *et al.* Genetic analysis of host resistance: Toll-like receptor signaling and immunity at large. *Annu. Rev. Immunol.* **24,** 353–389 (2006).
12. Winzler, C. *et al.* Maturation stages of mouse dendritic cells in growth factor-dependent long-term cultures. *J. Exp. Med.* **185,** 317–328 (1997).
13. Kawasaki, K., Gomi, K., Kawai, Y., Shiozaki, M. & Nishijima, M. Molecular basis for lipopolysaccharide mimetic action of Taxol and flavolipin. *J. Endotoxin Res.* **9,** 301–307 (2003).
14. Suzuki, K. G. *et al.* GPI-anchored receptor clusters transiently recruit Lyn and Gα for temporary cluster immobilization and Lyn activation: single-molecule tracking study 1. *J. Cell Biol.* **177,** 717–730 (2007).
15. Carpenter, G. & Ji, Q. Phospholipase C-γ as a signal-transducing element. *Exp. Cell Res.* **253,** 15–24 (1999).
16. Pugin, J. *et al.* Cell activation mediated by glycosylphosphatidylinositol-anchored or transmembrane forms of CD14. *Infect. Immun.* **66,** 1174–1180 (1998).
17. Kinoshita, T., Fujita, M. & Maeda, Y. Biosynthesis, remodelling and functions of mammalian GPI-anchored proteins: recent progress. *J. Biochem.* **144,** 287–294 (2008).
18. Moore, K. J. *et al.* Divergent response to LPS and bacteria in CD14-deficient murine macrophages. *J. Immunol.* **165,** 4272–4280 (2000).
19. Haziot, A. *et al.* Resistance to endotoxin shock and reduced dissemination of gram-negative bacteria in CD14-deficient mice. *Immunity* **4,** 407–414 (1996).
20. Jiang, Z. *et al.* CD14 is required for MyD88-independent LPS signaling. *Nature Immunol.* **6,** 565–570 (2005).
21. Kagan, J. C. *et al.* TRAM couples endocytosis of Toll-like receptor 4 to the induction of interferon-β. *Nature Immunol.* **9,** 361–368 (2008).
22. Shakhov, A. N., Collart, M. A., Vassalli, P., Nedospasov, S. A. & Jongeneel, C. V. κB-type enhancers are involved in lipopolysaccharide-mediated transcriptional activation of the tumor necrosis factor α gene in primary macrophages. *J. Exp. Med.* **171,** 35–47 (1990).
23. Dendorfer, U. Molecular biology of cytokines. *Artif. Organs* **20,** 437–444 (1996).
24. Aramburu, J. *et al.* Affinity-driven peptide selection of an NFAT inhibitor more selective than cyclosporin A. *Science* **285,** 2129–2133 (1999).
25. Schuh, K. *et al.* Retarded thymic involution and massive germinal center formation in NF-ATp-deficient mice. *Eur. J. Immunol.* **28,** 2456–2466 (1998).
26. Mahnke, K., Qian, Y., Knop, J. & Enk, A. H. Induction of $CD4^+/CD25^+$ regulatory T cells by targeting of antigens to immature dendritic cells. *Blood* **101,** 4862–4869 (2003).
27. Dobrovolskaia, M. A. & Vogel, S. N. Toll receptors, CD14, and macrophage activation and deactivation by LPS. *Microbes Infect.* **4,** 903–914 (2002).
28. Medzhitov, R. Origin and physiological roles of inflammation. *Nature* **454,** 428–435 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Akira, C. Kirschning and K. Miyake for mutant mice; M. Colonna for mutant mice and advice on PLC-γ2 activation; E. Serfling for mutant mice and advice on NFAT activation; A. Rao for the VIVIT peptide. We also thank Genopolis–Consorzio di Genomica Funzionale for microarray hybridization and data analysis and K. Mahnke for the anti-DEC205–OVA conjugate. This work was supported by grants from the CARIPLO Foundation, the European Commission 6th Framework Program (MUGEN and DC-THERA contracts), the European Commission 7th Framework Program (TOLERAGE and ENCITE contracts), the Associazione Italiana per la Ricerca sul Cancro (AIRC) and the and the Italian Ministry of Education and Research (COFIN).

Author Contributions F.G. conceived and oversaw the project and wrote the paper; I.Z. conceived the research and conducted most of the experiments with R.O.; M.C., M.C., M.R., G.C., B.C. and A.Z. helped with calcium experiments; G.C. and F.M. helped with experiments on D1 cells; M.F. helped with quantitative real-time PCR; A.E.R. performed the EMSA experiments; P.R.-C. provided advice.

Author Information All microarray data are available from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo) under accession code GSE15759. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.G. (francesca.granucci@unimib.it).

METHODS

Dendritic cells and macrophages. D1 cells and fresh bone marrow cells from C57BL/6 or mutant mice were cultured as previously described[29].

Mice and cells. C57BL/6 mice were purchased from Harlan. *Cd14*$^{-/-}$ mice were purchased from CNRS. OT-II mice were purchased from Charles River. *Tlr4*$^{-/-}$ mice were provided by S. Akira. *Plcg2*$^{-/-}$ 129×C57BL/6 and wild-type 129×C57BL/6 mice were provided by M. Colonna. TLR2/TLR4-double-deficient mice were provided by C. Kirschning. NFATc2-deficient mice were provided by E. Serfling, and MD2-deficient mice were provided by K. Miyake. All animals were housed under pathogen-free conditions. All experiments were carried out in accordance with the relevant laws and institutional guidelines.

Antibodies and chemicals. All the antibodies used for FACS analysis were purchased from BD Biosciences. TLR4-grade LPSs (*Escherichia coli*, 055:B5 and *Salmonella Typhimurium* S-form) were purchased from Alexis Corporation; non-TLR4-grade LPS (*E. coli* O55:B5) was purchased from Sigma. Recombinant soluble CD14 was purchased from CellSciences and Src inhibitor SU6656 from Sigma. Tacrolimus/FK506 (Fujisawa Pharmaceutical) was used at a concentration of 10 ng ml^{-1}.

Partial depletion and replenishment of cholesterol in the plasma membrane was performed according to ref. 14.

Indo1-AM and Fluo4-AM (Molecular Probes) were dissolved in DMSO. Stock solutions were diluted in Tyrode solution (154 mM NaCl, 4 mM KCl, 2 mM $CaCl_2$, 1 mM $MgCl_2$, 5 mM HEPES-NaOH, 5.5 mM D-glucose, adjusted to pH 7.35). The concentration of DMSO in the medium did not exceed 0.1%.

Western blot analysis. Wild-type and *Cd14*$^{-/-}$ BMDCs were stimulated for the indicated times with LPS (1 μg ml^{-1}) and lysed in the presence of Protein Inhibitor Cocktail (Sigma) and Phosphatase Inhibitor Cocktails (Sigma). Proteins from cell lysates were separated by standard SDS–PAGE and analysed by immunoblotting with antibodies specific for phosphorylated PLC-γ2 (from Cell Signaling) and for β-actin.

NFAT activation. Electrophoretic mobility shift assay (EMSA). ^{32}P-labelled DNA oligonucleotide probes were incubated for binding with 5 mg of nuclear extracts[30] for 20 min at room temperature in a buffer containing 5% glycerol, 50 mM NaCl, 20 mM Tris pH 7.9, 0.5 mM EDTA, 5 mM MgCl, 1 mM dithiothreitol, 100 ng ml^{-1} poly(dI-dC), and 50 ng ml^{-1} bovine serum albumin (BSA) in a final volume of 15 ml. The protein–DNA complexes were then separated on a 5% polyacrylamide gel (29:1 acrylamide/bisacrylamide ratio) and were visualized by autoradiography. The sequences of the oligonucleotide probes (sense strand) used in the gel shift assay were as follows: NFATc consensus site, 5′-GCCCAAAGAGGAAAATTTGTTTCATACAG-3′; NFATc mutant site, 5′-GCCCAAAGcttAAAATTTGTTTCATACAG-3′; Oct consensus site, 5′-TCGTCAAACTCTGCTAATTAGCAATGCTGAGAAA-3′.

Immunocytochemistry. Cells were fixed in formaldehyde and spun onto glass slides (2×10^5 cells per slide in 200 μl) at 300 r.p.m. for 5 min. Cells were then permeabilized in chilled methanol and washed extensively with PBS. Fc block antibody (CD16/CD32, BD Pharmingen), rabbit anti-mouse NFATc2 (1:250, ImmunoGlobe), rabbit anti-mouse NFATc3 (1:250, Santacruz) and donkey anti-rabbit Alexa Fluor 555 antibodies (1:500, Molecular Probes) diluted in PBS plus 0.1% BSA were added and incubated at room temperature. DAPI (1:1,000, Sigma Aldrich) was added and the samples incubated for 5 min at room temperature.

IL-2, TNF-α and IL-6 measurements. ELISAs were performed with the DuoSet kits (R & D).

Plasmids and retroviral infection of dendritic cells. VIVIT–GFP DNA was inserted into the PINCO retroviral vector[31], using the HindIII and NotI restriction sites. Standard molecular biology methods were used for cDNA cloning. All the reagents used were from Invitrogen.

The production of high-titre vectors and the D1 infection protocol have been described elsewhere[31]. Transduction efficiency was evaluated by FACS analysis and was generally around 30%. GFP$^+$ cells were sorted and cultured as previously described[31].

***In vivo* production of IL-2 and TNF-α by spleen cells.** Aliquots of 10 μg of LPS or 10 μg of LPS and 5 mM Tpg (in a final volume of 50 μl) were injected into mouse spleens. After 3 h, spleens were removed and single-cell suspensions prepared. The cells suspensions were incubated with brefeldin A (10 μg ml^{-1}, Sigma-Aldrich) for 5 h. The cells were fixed using BD Phosflow Lyse/Fix buffer (BD Biosciences), permeabilized with BD Phosflow Perm Buffer III (BD Biosciences) and stained with PE-labelled anti-CD11c, APC-labelled anti-IL-2 and anti-TNF-α monoclonal antibodies or APC-labelled isotype control antibody. Cells were then analysed on a FACScalibur (Becton Dickinson).

Microarray experiment. Total RNA was extracted by the double extraction protocol: RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction (Trizol Invitrogen) followed by a Qiagen RNeasy clean-up procedure. Total RNA integrity was assessed with a Agilent Bioanalyser and the RNA Integrity Number (RIN) was calculated. Only high-quality RNA preparations, with RIN greater than 8.5, were used for microarray analysis.

Three micrograms of total RNA was used for cRNA target preparation according to the Affymetrix GeneChip Expression Analysis Technical Manual (Affymetrix) using the one cycle target labelling kit and according to the manufacturer's instructions. Ten micrograms of biotinylated cRNA was hybridized to the Affymetryx GeneChip Mouse Genome 430A 2.0 arrays.

Bioconductor[32] was used for most data handling. The Guanine Cytosine Robust Multi-array Analysis GCRMA[33] method was used to calculate probe set intensity. The normalization method applied was the quantile.

***In vitro* survival assay.** Cells were seeded in 96-well plates at a concentration of 0.5×10^6 cells ml^{-1} in 100 μl of complete IMDM plus GM-CSF (for BMDCs) or M-CSF (for bone-marrow macrophages) in the presence or absence of LPS (1 μg ml^{-1}) and/or thapsigargin (50 nM). After 24 h, 100 μl of complete IMDM was added to the culture. Survival of cells was measured using the CellTiter-Blue Cell Viability Assay (Promega).

***In vivo* survival assay.** Four-to-six-week-old mice, showing an average of 90 million total spleen cells, were injected intravenously with 1 μg g^{-1} of LPS. At various times thereafter, spleens were collected, single-cell suspensions produced and total spleen cells counted. Absolute numbers of dendritic cells were determined by FACS analysis using CD11c–APC-conjugated, CD8α–PE-conjugated, and CD11b–FITC-conjugated antibodies.

T-cell activation. Mice were injected intravenously with LPS (1 μg g^{-1}) and anti-DEC205–OVA complex (10 μg) and 8 h later received CFSE-labelled CD4$^+$ T cells (4×10^6) from OT-II mice. T-cell proliferation was measured in the spleen after 48 h. One week after immunization total CD4$^+$ T cells were recovered and re-stimulated *in vitro* with OVA peptide in the presence of splenic APCs. IFN-γ production was measured by ELISA after 48 h of culture.

Real-time quantitative PCR. Total RNA was extracted from 10^6 cells using the TRIZOL reagent according to the recommended procedure (Gibco-BRL). Single-strand cDNA was synthesized using High-capacity cDNA Reverse Transcription Kits (Perkin-Elmer, Applied Biosystem Division). The NanoDrop (TermoScientific) was used to titre mRNA and amplification was performed using the Power Sybr Green PCR Master Mix (Perkin Elmer).

Statistical analysis. Means were compared by paired or unpaired *t*-tests. Data are expressed and plotted as means ± s.d. values. Statistical significance was defined as $P < 0.05$. Sample sizes for each experimental condition are provided in the figures and the respective legends.

29. Granucci, F. *et al.* Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. *Nature Immunol.* **2,** 882–888 (2001).
30. Schreiber, E., Matthias, P., Muller, M. M. & Schaffner, W. Rapid detection of octamer binding proteins with 'mini-extracts', prepared from a small number of cells. *Nucleic Acids Res.* **17,** 6419 (1989).
31. Granucci, F. *et al.* The scavenger receptor MARCO mediates cytoskeleton rearrangements in dendritic cells and microglia. *Blood* **102,** 2940–2947 (2003).
32. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264 (2003).
33. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).

LETTERS

T cells dampen innate immune responses through inhibition of NLRP1 and NLRP3 inflammasomes

Greta Guarda[1], Catherine Dostert[1], Francesco Staehli[1], Katrin Cabalzar[1], Rosa Castillo[1], Aubry Tardivel[1], Pascal Schneider[1] & Jürg Tschopp[1]

Inflammation is a protective attempt by the host to remove injurious stimuli and initiate the tissue healing process[1]. The inflammatory response must be actively terminated, however, because failure to do so can result in 'bystander' damage to tissues and diseases such as arthritis or type-2 diabetes. Yet the mechanisms controlling excessive inflammatory responses are still poorly understood. Here we show that mouse effector and memory $CD4^+$ T cells abolish macrophage inflammasome-mediated caspase-1 activation and subsequent interleukin 1β release in a cognate manner. Inflammasome inhibition is observed for all tested NLRP1 (commonly called NALP1) and NLRP3 (NALP3 or cryopyrin) activators, whereas NLRC4 (IPAF) inflammasome function and release of other inflammatory mediators such as CXCL2, interleukin 6 and tumour necrosis factor are not affected. Suppression of the NLRP3 inflammasome requires cell-to-cell contact and can be mimicked by macrophage stimulation with selected ligands of the tumour necrosis factor family, such as CD40L (also known as CD40LG). In a NLRP3-dependent peritonitis model, effector $CD4^+$ T cells are responsible for decreasing neutrophil recruitment in an antigen-dependent manner. Our findings reveal an unexpected mechanism of inflammasome inhibition, whereby effector and memory T cells suppress potentially damaging inflammation, yet leave the primary inflammatory response, crucial for the onset of immunity, intact.

A tight control of innate immunity is essential because morbidity from infectious diseases can be caused directly by the pathogen, as well as by a disproportionate immune response. Recent studies delineated a direct role for T cells not only in pathogen clearance but also in regulating adaptive and innate immunity[2–6]. Regulatory T cells (Tregs) were found to suppress innate-immunity-driven inflammation[2–4]. Similarly, non-Treg T cells were reported to reduce production of interferon (IFN)-γ and tumour necrosis factor (TNF) by innate immune cells in an antigen-independent manner[5,6].

Interleukin (IL)-1β (also known as IL1B) is a major pro-inflammatory cytokine. Conversion of the proIL-1β precursor to the active protein (p17) requires the assembly of intracellular complexes called inflammasomes[7], which include proteins of the Nod-like receptor (NLR) family. Upon detection of pathogen-associated or danger-associated molecular patterns, inflammasomes recruit and activate caspase-1, which in turn proteolytically processes proIL-1β (refs 8, 9).

Considering the strong pro-inflammatory activity of IL-1β, its processing and secretion must be tightly regulated. We therefore explored the possibility that Tregs might affect inflammasome function. Tregs and, as controls, naive and memory $CD4^+$ and $CD8^+$ T cells were purified from splenocytes and incubated with bone-marrow-derived macrophages (BMDMs) in the presence of soluble anti-CD3 antibody. After overnight co-culture, lipopolysaccharide (LPS) was added to induce proIL-1β synthesis by BMDMs, followed by ATP stimulation to activate the NLRP3 inflammasome, IL-1β processing and release[9]. Surprisingly, secretion of mature IL-1β was blocked in the presence of $CD4^+$ memory T cells, whereas it was only marginally altered by other T-cell subsets (Fig. 1a and Supplementary Fig. 1a). In line with this reduction, caspase-1 processing in BMDMs was abrogated in the presence of $CD4^+$ memory T cells (Fig. 1b). IL-1β and cleaved caspase-1 originated from BMDMs and not T cells, as shown in Supplementary Fig. 2. Furthermore, decreased IL-1β release was not a consequence of diminished proIL-1β or impaired BMDM viability (Fig. 1b and Supplementary Fig. 3a). The inhibitory capacity was common to all the major memory $CD4^+$ T-cell subsets (Supplementary Fig. 1b, c)[10,11].

We next asked whether T-cell receptor (TCR) triggering was needed for suppression of inflammasome activation by memory $CD4^+$ T lymphocytes and found that the inhibition was completely dependent on the presence of anti-CD3 antibody (Fig. 1c). In addition to *ex vivo* memory $CD4^+$ T lymphocytes, *in-vitro*-activated $CD4^+$ (primed with plastic-bound anti-CD3 and soluble anti-CD28) and, to a lesser extent, *in-vitro*-activated $CD8^+$ T cells also blocked BMDM inflammasome activation when restimulated in the presence of soluble anti-CD3 (Supplementary Fig. 4a, b). TCR transgenic effector OT-II T cells, which can be restimulated with the cognate ovalbumin ($OVA_{323–339}$) peptide, suppressed caspase-1 activation as efficiently as observed with anti-CD3 restimulation (Fig. 1d and Supplementary Fig. 5b, c). Moreover, $CD4^+$ effector T cells also inhibited IL-1β release by bone-marrow-derived dendritic cells (BMDCs) indicating that their suppressive activity is not specific to BMDMs (Fig. 1e). Taken together, these results show that *ex vivo* memory and *in-vitro*-activated effector $CD4^+$ T cells can block inflammasome function and that antigenic restimulation is an essential prerequisite for this inhibition.

We next tested whether this suppression applies to other NLRP3 activators and found that $CD4^+$ T cells also diminished BMDM caspase-1 activation and IL-1β release in response to monosodium urate crystals (MSU), asbestos, alum or nigericin (Fig. 2a, b)[9]. The less efficient inhibition observed on nigericin stimulation is explained by the potency of the toxin at the applied dose (Supplementary Fig. 5).

To understand whether the inhibition exerted by effector and memory $CD4^+$ T cells on BMDMs was restricted to the pro-inflammatory cytokines dependent on caspase-1, a more comprehensive analysis of the effect of $CD4^+$ T cells on the production of inflammatory mediators by BMDMs was performed. As expected, secretion of caspase-1-dependent IL-18 was strongly impaired (Fig. 2c). In contrast, LPS-induced CXCL2, IL-12, TNF and IL-6 levels were either unaltered or enhanced in the presence of $CD4^+$ T cells, as observed also for untreated or CpG-treated BMDMs (Fig. 2c and

[1]Department of Biochemistry, University of Lausanne, Chemin des Boveresses 155, CH-1066 Epalinges, Switzerland.

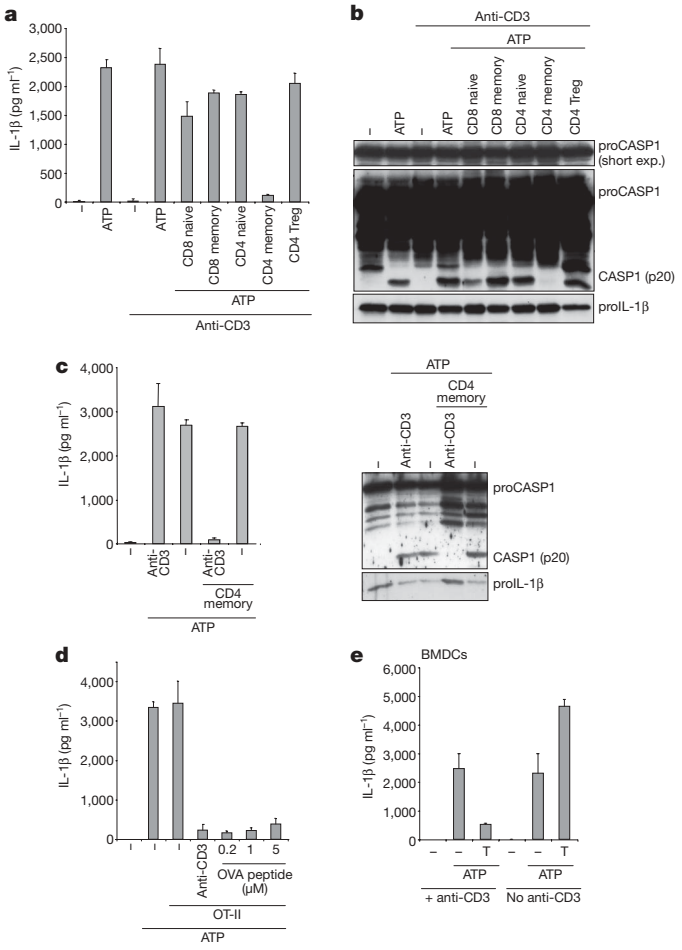


Figure 1 | Memory and effector T cells inhibit the BMDM NLRP3 inflammasome in an antigen-dependent manner. a, **b**, FACS-sorted naive, memory and regulatory $CD4^+$ and naive and memory $CD8^+$ T cells were cultured overnight together with BMDMs in the presence of anti-CD3 antibody. Thereafter, BMDMs were primed with LPS for 4 h and then stimulated with ATP. IL-1β release was assessed by ELISA (**a**), and caspase-1 (CASP1) activation in cell lysates was assessed by western blot (**b**). exp., exposure. **c**, ATP-induced IL-1β release and caspase-1 activation by BMDMs cultured overnight in the presence of FACS-sorted memory $CD4^+$ T cells with or without anti-CD3 were measured in supernatants and cell extracts. **d**, ATP-induced IL-1β release by BMDMs cultured overnight with *in-vitro*-activated OT-II cells restimulated by anti-CD3 or $OVA_{323–339}$ peptide. **e**, ATP-induced IL-1β release by BMDCs cultured overnight in the presence of *in-vitro*-primed effector $CD4^+$ T cells. **a**, **c**–**e**, ELISA data represent mean and s.d. of three individual experimental points.

Supplementary Fig. 6). When cultured together with effector $CD4^+$ T lymphocytes, BMDMs also showed augmented granularity (Supplementary Fig. 3b), increased expression of CD86, major histocompatibility complex (MHC) class II and CD40 (Supplementary Fig. 6). Thus, effector $CD4^+$ T cells appear to promote general BMDM activation, while exclusively blocking the caspase-1 axis.

To understand whether $CD4^+$ T cells solely affect the NLRP3 inflammasome, we tested their effect on NLRC4 and NLRP1 inflammasomes by infecting BMDMs with *Salmonella typhimurium* or stimulating with *Bacillus anthracis* lethal toxin, respectively[12]. Although caspase-1 activation and IL-1β secretion upon *S. typhimurium* infection were not affected (Fig. 2d and Supplementary Fig. 7), NLRP1 inflammasome activation by lethal toxin was inhibited (Fig. 2e), indicating differential modes of action among members of the NLR family.

To investigate whether abrogation of caspase-1 activity was mediated by a soluble factor or required cell-to-cell contact, BMDMs were incubated with conditioned media from overnight T-cell–BMDM co-culture, and stimulated with alum (Fig. 3a). This failed to block IL-1β release, indicating that inhibition is dependent on direct T-cell–BMDM contact and consistent with the observation that the minimal T-cell-to-BMDM ratio required for complete inhibition is around 1:1 (Supplementary Fig. 8a). We next investigated the kinetics of inflammasome inhibition. Caspase-1 cleavage in BMDMs was maximally suppressed eight hours after co-culture with effector $CD4^+$ T cells in the presence of anti-CD3 (Fig. 3b). Given the dynamic nature of T-cell–antigen-presenting cell (APC) interactions *in vivo*[13], we also assessed the kinetics of inflammasome inhibition using effector T cells pre-restimulated with plastic-bound anti-CD3 and soluble anti-CD28. When using pre-reactivated T cells, the inflammasome was completely inhibited within three hours (Fig. 3c and Supplementary Fig. 8b). Moreover, inhibition was also observed with fixed pre-reactivated T cells, in agreement with the suggested membrane-bound nature of the inhibitory factor (Fig. 3c and Supplementary Fig. 8b).

Recent evidence implicates CD200 or IFN-γ in downregulating excessive inflammatory responses[14,15]. However, incubation of BMDMs with CD200-fragment crystallizable (Fc) fusion protein or IFN-γ did not affect inflammasome activity (Supplementary Fig. 9). We thus turned our interest to TNF family ligands, known to have a central role in the T-cell–APC cross-talk and for which expression is increased upon T-cell stimulation[16–19] (Supplementary Fig. 10). Because our results indicate that the inflammasome-inhibiting factor is membrane-bound, we used hexameric-TNF ligand constructs, which mimic the activity of membrane-bound ligands[20,21]. Indeed, among the tested ligands, Fc–RANKL (also known as TNFSF11; receptor activator of nuclear factor kappa B ligand), Fc–CD40L, Fc–LIGHT (TNFSF14; lymphotoxin homologue inducible competing with glycoprotein D for HVEM (expressed on T cells)), Fc–41BBL (TNFSF9), Fc–CD30L (TNFSF8), Fc–OX40L (TNFSF4) and, to a lesser extent, Fc–TNF diminished IL-1β secretion upon ATP stimulation (Fig. 3d). For a more detailed analysis, we focused on CD40L, the expression of which is associated with $CD4^+$ T-cell activation[18,19]. A highly active CD40L construct, the adiponectin (also known as ACRP) fusion protein ADIPOQ–CD40L, suppressed alum-induced IL-1β release and caspase-1 activation in a dose-, CD40- and time-dependent manner, without affecting BMDM viability (Fig. 3e, f and Supplementary Fig. 11a, b). As observed for $CD4^+$ T cells, ADIPOQ–CD40L effectively suppressed the inflammasome function triggered by all tested NLRP3 activators, whereas the secretion of caspase-1-independent inflammatory mediators was unaltered or even enhanced (Supplementary Fig. 11c, d).

To investigate the importance of inflammasome suppression *in vivo*, we set up an alum-induced, NLRP3-dependent peritonitis model in which neutrophil recruitment was measured in the context of a primary or secondary antigenic challenge[22]. Mice were either immunized with ovalbumin and CpG-DNA intraperitoneally (i.p.) or injected with CpG-DNA alone, as a control. Ten days later, peritonitis was induced by i.p. injection of the NLRP3 activator alum as a control, or alum together with ovalbumin, to restimulate effector T cells. Neutrophil recruitment to the inflamed peritoneum was subsequently assessed (Fig. 4a). Neutrophil numbers were significantly reduced only in ovalbumin-immunized mice, to which alum was coinjected with ovalbumin, indicating that the suppression is caused by the adaptive immune response. Similar results were observed for eosinophil influx (Supplementary Fig. 12). We next repeated the assay in $Igh^{-/-}$ mice, which lack B cells (Fig. 4b). B-cell deficiency did not affect the suppression of neutrophil recruitment, indicating a key role for T cells in inhibiting neutrophil influx. This was confirmed using $Tcrbd^{-/-}$ mice, which lack T cells, in which neutrophil numbers were not decreased in immunized mice (Fig. 4c). Finally, when the assay was repeated in $Cd4^{-/-}$ mice (Fig. 4d), no reduction was observed, demonstrating that the $CD4^+$ T-cell response has a predominant role in suppressing NLRP3-dependent neutrophil influx.

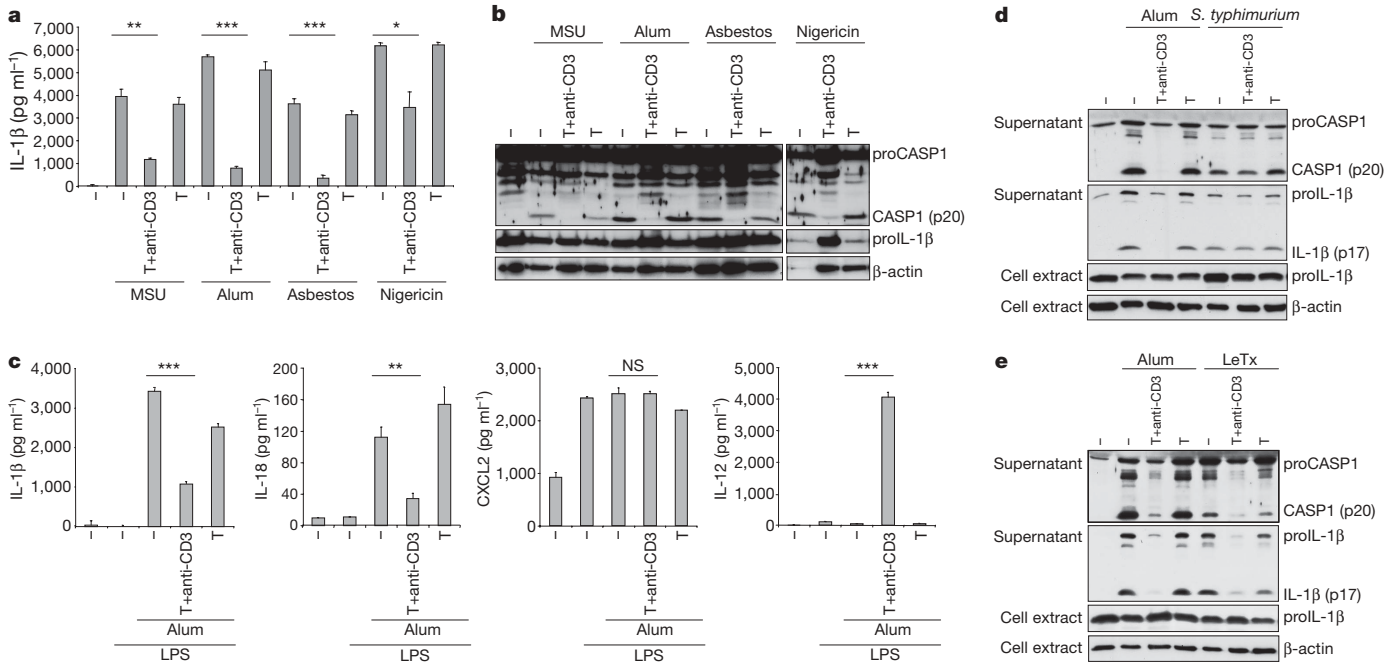


Figure 2 | T cells selectively inhibit NLRP3 and NLRP1 inflammasomes. BMDMs were cultured overnight with *in-vitro*-activated effector $CD4^+$ T cells in the presence or absence of anti-CD3 antibody. LPS was added to prime the BMDMs. **a**, **b**, IL-1β release measured by ELISA (**a**) and caspase-1 activation in cell lysates (**b**) after stimulation with MSU, alum, asbestos or nigericin. **c**, IL-1β, IL-18, CXCL2 and IL-12 release by LPS-primed BMDMs was measured by ELISA after alum stimulation. **d**, **e**, Caspase-1 and IL-1β cleavage and secretion were assessed by western blot after *S. typhimurium* infection (**d**) or lethal toxin (LeTx) stimulation (**e**). **a**, **c**, Data represent mean ± s.d. of three individual experimental points. $*P \leq 0.05$; $**P \leq 0.01$; $***P \leq 0.001$; NS, not significant.

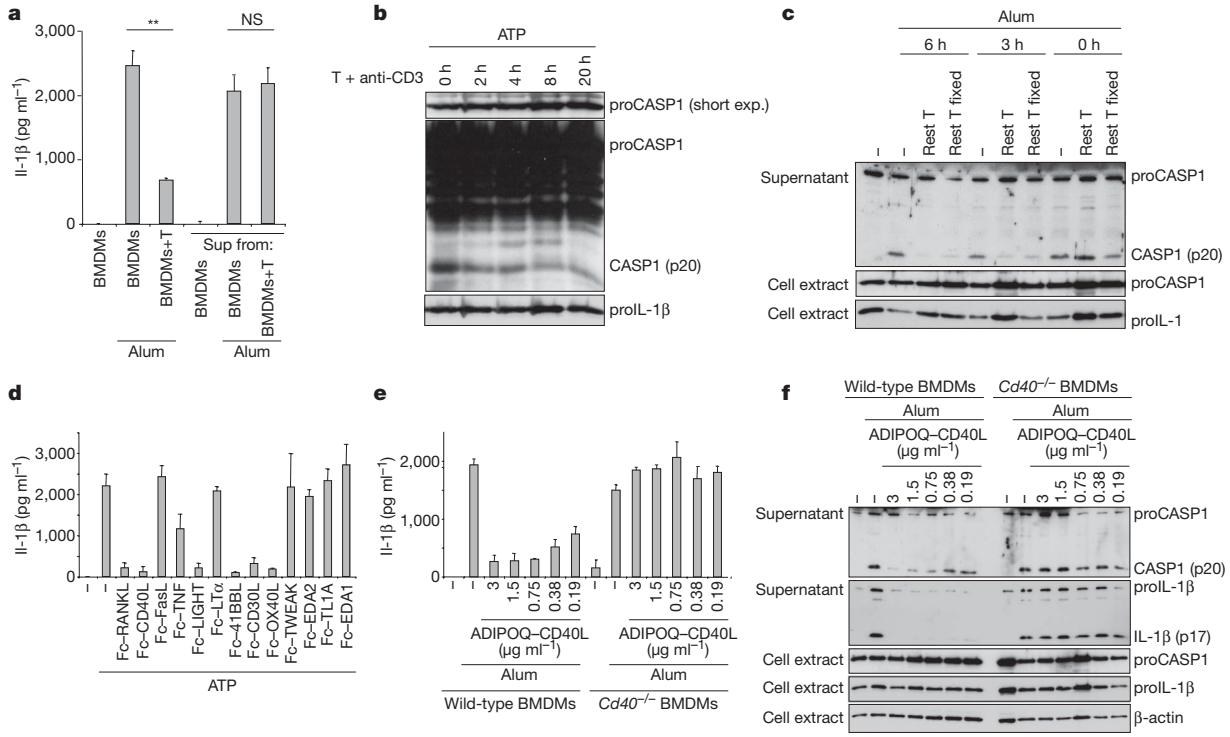


Figure 3 | TNF family ligands expressed by T cells inhibit the NLRP3 inflammasome. **a**, Supernatant (Sup) from overnight T-cell–BMDM cultures in the presence of anti-CD3 or control supernatant from BMDMs alone were added to BMDMs. After 8 h, LPS-primed BMDMs were treated with alum and IL-1β release was measured. **b**, T cells and soluble anti-CD3 were added to BMDMs for the indicated times before ATP stimulation. LPS was added 4 h before ATP stimulation. Caspase-1 activation in cell lysates was assessed by western blot. **c**, Caspase-1 activation in LPS-primed BMDMs cultured together with pre-restimulated (rest) (glutaraldehyde fixed or not) *in-vitro*-generated effector T cells was assessed by western blot. **d**, BMDMs incubated overnight with the indicated Fc–TNF family ligands were primed with LPS and stimulated with ATP. IL-1β release was quantified by ELISA. **e**, **f**, Wild-type and *Cd40*$^{-/-}$ BMDMs were incubated overnight with the indicated doses of ADIPOQ–CD40L, LPS-primed and stimulated with alum. IL-1β release was measured by ELISA (**e**) and caspase-1 cleavage was assessed by western blot (**f**). **a**, **d**, **e**, ELISA data represent mean ± s.d. of three individual experimental points. $**P \leq 0.01$; NS, not significant.

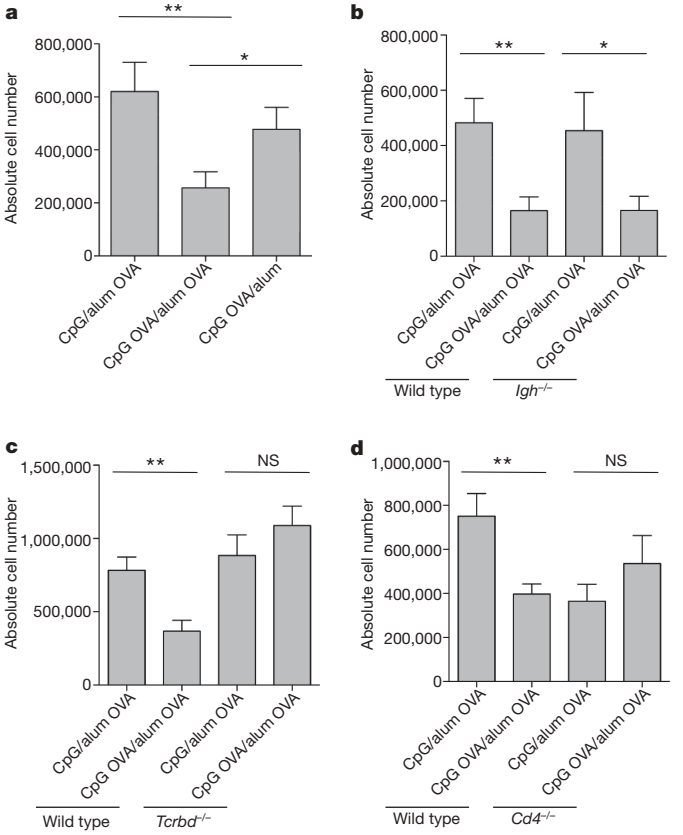


Figure 4 | Neutrophil recruitment to the peritoneum is suppressed in mice with effector T cells. a, Absolute numbers of neutrophils recruited to the peritoneum of C57BL/6 mice treated with CpG-DNA alone and rechallenged i.p. on day 10 with alum and ovalbumin (CpG/alum OVA), mice immunized with CpG-DNA and ovalbumin and rechallenged with alum and ovalbumin (CpG OVA/alum OVA) or mice immunized with CpG-DNA and ovalbumin and rechallenged with alum alone (CpG OVA/alum). The peritoneum was lavaged 12–14 h later. **b**–**d**, The two first conditions described in **a** were repeated, except that wild-type C57BL/6 and $Igh^{-/-}$ (**b**), $Tcrbd^{-/-}$ (**c**) or $Cd4^{-/-}$ (**d**) were used. Values are the mean ± s.e.m. of 6–8 mice per group (**a**–**d**). $*P \leq 0.05$; $**P \leq 0.01$; NS, not significant.

Our data support a mechanism whereby $CD4^+$ effector and memory T cells selectively block NLRP3 and NLRP1 inflammasomes in an antigen-dependent manner. However, caspase-1-independent inflammatory mediators were not affected or even enhanced, in agreement with the well-established role for T lymphocytes in positively regulating APC activation[18,23]. T-cell activation and antigenic stimulation were found to be essential prerequisites for the blockade of the NLRP inflammasomes, consistent with the preferential migration of activated T cells to inflammatory sites[11]. Together, our data suggest that lymphocyte infiltration acts as a feedback loop, whereby effector or memory T cells edit the quality of the inflammatory mediators produced during the late phases of primary responses or on secondary challenges. The selective regulation of potent inflammatory cytokines such as IL-1β and IL-18 is likely to be a mechanism by which innate immunity-driven inflammation is controlled, yet the full competence in antigen presentation is maintained to guarantee efficient T-cell restimulation.

The TNF superfamily is strongly implicated in the cross-talk between T cells and APCs[16–18]. The finding that this family has the ability to turn off the inflammasome is consistent with its established role in shaping the adaptive immune response by acting on innate immune cells[18]. It is tempting to speculate that the TNF family might switch inflammation from a broad response, comprising cytokines with highly pleiotropic activity such as IL-1β, to a more directed response, favouring the production of cytokines with more specific effects.

Uncontrolled IL-1β activity causes considerable collateral damage, as seen in patients with a hyperactive inflammasome[24]. A better understanding of the molecular mechanisms of inflammasome inhibition may thus provide new opportunities for the development of anti-inflammatory drugs and shed new light on inflammatory disorders that are currently poorly understood.

METHODS SUMMARY

Mice. All transgenic and knockout mouse strains used have been described elsewhere. Mice were treated in accordance with the Swiss Federal Veterinary Office guidelines.

TNF family ligands. Recombinant murine TNF family ligands (RANKL, CD40L, Fasl, TNF, LIGHT, lymphotoxin α (LTα, also known as LTA), 41BBL, CD30L, OX40L, TNF family member with weak apoptosis-inducing activity (TWEAK, also known as TNFSF12), TNF family ligand 1A (TL1A, also known as TNFSF15), ectodysplasin A (EDA) 1 and 2) were expressed in supernatants of transfected 293T cells as Fc-fusion proteins (concentrations of about 1–10 $\mu g\,ml^{-1}$)[20]. Mouse ADIPOQ–CD40L was purchased from Adipogen.

***In vitro* stimulation experiments.** A total of $1.6–1.8 \times 10^5$ (unless otherwise indicated) *ex vivo* fluorescence-activated cell sorting (FACS)-sorted or *in-vitro*-activated T cells were cultured with 5×10^4 differentiated BMDMs or BMDCs overnight (14–16 h) in the presence of 0.5 $\mu g\,ml^{-1}$ soluble anti-CD3 or the indicated dose of $OVA_{323–339}$ peptide (Anaspec). Alternatively, T cells were replaced by TNF family ligands. Fc-fusion TNF ligands were used in the presence of 1 $\mu g\,ml^{-1}$ protein A (Sigma). The next morning, BMDMs and BMDCs were primed for 4 h with 10 $ng\,ml^{-1}$ ultrapure LPS (Invivogen). Next, stimulations with inflammasome activators were carried out. ATP (500 μM), nigericin (6 μM) and MSU crystals (150 $\mu g\,ml^{-1}$)[25] were from Sigma. Asbestos (100 $\mu g\,ml^{-1}$) was from SPI-CHEM and alum (300 $\mu g\,ml^{-1}$) was from Pierce Biochemicals (Imject-alum). ATP stimulation was performed for 45 min, and other stimulations for 150 min (unless otherwise indicated). For *S. typhimurium* infection, cell cultures were infected by centrifugation with a multiplicity of infection (MOI) of 90, followed by 20 min at 37 °C. Medium was then substituted with gentamycin-complemented (50 $\mu g\,ml^{-1}$) medium and culture was continued for 3 h at 37 °C. *B. anthracis* lethal toxin, consisting of protective antigen and lethal factor, was from List Biological Laboratories; stimulation with 5 $\mu g\,ml^{-1}$ lethal toxin was performed on cells of Balb/c origin for 4 h.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 March; accepted 16 April 2009.
Published online 3 June 2009.

1. Medzhitov, R. Origin and physiological roles of inflammation. *Nature* **454,** 428–435 (2008).
2. Miyara, M. & Sakaguchi, S. Natural regulatory T cells: mechanisms of suppression. *Trends Mol. Med.* **13,** 108–116 (2007).
3. Maloy, K. J., Antonelli, L. R., Lefevre, M. & Powrie, F. Cure of innate intestinal immune pathology by $CD4^+CD25^+$ regulatory T cells. *Immunol. Lett.* **97,** 189–192 (2005).
4. Garrett, W. S. *et al.* Communicable ulcerative colitis induced by T-bet deficiency in the innate immune system. *Cell* **131,** 33–45 (2007).
5. Kim, K. D. *et al.* Adaptive immune cells temper initial innate responses. *Nature Med.* **13,** 1248–1252 (2007).
6. Palm, N. W. & Medzhitov, R. Not so fast: adaptive suppression of innate immunity. *Nature Med.* **13,** 1142–1144 (2007).
7. Martinon, F., Burns, K. & Tschopp, J. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-beta. *Mol. Cell* **10,** 417–426 (2002).
8. Martinon, F., Gaide, O., Petrilli, V., Mayor, A. & Tschopp, J. NALP inflammasomes: a central role in innate immunity. *Semin. Immunopathol.* **29,** 213–229 (2007).
9. Yu, H. B. & Finlay, B. B. The caspase-1 inflammasome: a pilot of innate immune responses. *Cell Host Microbe* **4,** 198–208 (2008).
10. Sallusto, F., Lenig, D., Forster, R., Lipp, M. & Lanzavecchia, A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature* **401,** 708–712 (1999).
11. Bromley, S. K., Mempel, T. R. & Luster, A. D. Orchestrating the orchestrators: chemokines in control of T cell traffic. *Nature Immunol.* **9,** 970–980 (2008).
12. Boyden, E. D. & Dietrich, W. F. NALP1/NLRP1b controls mouse macrophage susceptibility to anthrax lethal toxin. *Nature Genet.* **38,** 240–244 (2006).
13. Mempel, T. R., Henrickson, S. E. & Von Andrian, U. H. T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. *Nature* **427,** 154–159 (2004).
14. Snelgrove, R. J. *et al.* A critical function for CD200 in lung immune homeostasis and the severity of influenza infection. *Nature Immunol.* **9,** 1074–1083 (2008).

15. Sun, K. & Metzger, D. W. Inhibition of pulmonary antibacterial defense by interferon-gamma during recovery from influenza infection. *Nature Med.* **14,** 558–564 (2008).
16. Croft, M. Co-stimulatory members of the TNFR family: keys to effective T-cell immunity? *Nature Rev. Immunol.* **3,** 609–620 (2003).
17. Watts, T. H. TNF/TNFR family members in costimulation of T cell responses. *Annu. Rev. Immunol.* **23,** 23–68 (2005).
18. O'Sullivan, B. & Thomas, R. CD40 and dendritic cell function. *Crit. Rev. Immunol.* **23,** 83–107 (2003).
19. Martin-Fontecha, A. *et al.* $CD40L^+$ $CD4^+$ memory T cells migrate in a CD62P-dependent fashion into reactive lymph nodes and license dendritic cells for T cell priming. *J. Exp. Med.* **205,** 2561–2574 (2008).
20. Bossen, C. *et al.* Interactions of tumor necrosis factor (TNF) and TNF receptor family members in the mouse and human. *J. Biol. Chem.* **281,** 13964–13971 (2006).
21. Holler, N. *et al.* Two adjacent trimeric Fas ligands are required for Fas signaling and formation of a death-inducing signaling complex. *Mol. Cell. Biol.* **23,** 1428–1440 (2003).
22. Kool, M. *et al.* Cutting edge: alum adjuvant stimulates inflammatory dendritic cells through activation of the NALP3 inflammasome. *J. Immunol.* **181,** 3755–3759 (2008).
23. Burger, D. & Dayer, J. M. The role of human T-lymphocyte-monocyte contact in inflammation and tissue destruction. *Arthritis Res.* **4** (Suppl 3), S169–S176 (2002).
24. McDermott, M. F. & Tschopp, J. From inflammasomes to fevers, crystals and hypertension: how basic research explains inflammatory diseases. *Trends Mol. Med.* **13,** 381–388 (2007).
25. Martinon, F., Petrilli, V., Mayor, A., Tardivel, A. & Tschopp, J. Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* **440,** 237–241 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Acha-Orbea, P. Launois, E. Luescher, H. R. MacDonald, P. Romero, M. Kopf and R. Flavell for knockout and transgenic mice; H. Acha-Orbea and M. Eckert for critical reading and comments; A. Yazdi for discussions; and M. Braun and D. Labes for technical help. This work was in part supported by the Swiss National Science foundation and the NCCR Molecular Oncology. C.D. is supported by an EMBO long-term fellowship.

Author Contributions G.G. performed and designed the experiments and wrote the paper. C.D., F.S., K.C., R.C. and A.T. contributed to performing the experiments. P.S. provided the TNF ligand reagents and discussed the data throughout the experimental phase. J.T. designed and discussed the experiments and supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.T. (jurg.tschopp@unil.ch).

METHODS

Mice. OT-II TCR transgenic, *Tcrβδ*$^{-/-}$, *Cd4*$^{-/-}$ (all on C57BL/6 background) mice were originally obtained from Jackson Laboratory. *Foxp3*–enhanced green fluorescent protein knock-in[26] mice were provided by B. Malissen. *Cd40*$^{-/-}$[27], caspase-1$^{-/-}$[28], *Igh*$^{-/-}$[29] (all on C57BL/6 background) mice were bred at the University of Lausanne. Six- to ten-week-old C57BL/6 and Balb/c mice were purchased from Harlan. For all the experiments shown, T cells and BMDMs on C57BL/6 background were used, with the exception of the *B. anthracis* lethal toxin stimulation, which was performed with cells on the Balb/c background[12]. Mice were treated in accordance with the Swiss Federal Veterinary Office guidelines.

FACS sorting and analysis. Naive, memory and regulatory CD4$^+$ cells and naive and memory CD8$^+$ cells were sorted from the spleen by cell sorting (FACSAria, BD Biosciences) as CD4$^+$ CD44^{low} CD25$^-$, CD4$^+$ CD44^{high} CD25$^-$, CD4$^+$ CD44$^{intermediate–high}$ CD25$^+$, CD8$^+$ CD44^{low} CD25$^-$ and CD8$^+$ CD44^{high} CD25$^-$, respectively, using specific antibodies (FITC-labelled anti-CD25 (7D4) from BD Biosciences, PE-labelled anti-CD44 (IM7), APC-labelled anti-CD4 (L3T4) and APC-labelled anti-CD8 (Ly-2), all from eBioscience). Before sorting, CD4$^+$ and CD8$^+$ T cells were enriched using anti-CD4 magnetic beads (Miltenyi) and APC-labelled anti-CD8 followed by anti-APC magnetic beads. The following monoclonal antibodies were used: anti-CD16/32 (93), anti-CD44 (IM7), anti-CD4 (L3T4), anti-CD8α (Ly-2), anti-CD3 (145-2C11), anti-CD25 (PC61.5), anti-CD11b (M1/70), anti-F4/80 (BM8), anti-CD86 (GL1), anti-MHCII (M5/114.15.2), anti-CCR7 (4B12), anti-CD62L (MEL14), anti-CD40L (MR1), anti-CD40 (1C10) and anti-RANKL (IK22/5) from eBioscience, anti-Ly-6C (AL-21) and anti-Ly-6G (1A8) from BD Biosciences and anti-CXCR3 (220803) from R&D Systems. CD30–Fc was used to detect CD30L. Cell surface four/six colour stainings were performed with the appropriate combinations of FITC-, PE-, PerCP-, PECy7-, APC-, APC-Cy7- or biotin-conjugated antibodies and APC-Cy7-labelled streptavidin (eBioscience). PI, used to exclude dead cells, was from Sigma. For intracellular staining, brefeldin A (Axxora) was added to a final concentration of 10 μg ml^{-1}. After 4 h, cells were collected with 6 mM EDTA in PBS and fixed with 4% formaldehyde. Cells were then permeabilized with saponin (Sigma) solution (0.5% saponin in PBS) and stained with anti-IL-6 (MP5-20F3, from BD Biosciences) and anti-TNF (MP6-XT22, from eBioscience) antibodies. To gate on BMDMs only, T cells were either stained with anti-CD4 antibody or labelled with CFSE before the co-culture and excluded from the analysis. Data were acquired on a FACSCalibur or a FACSCanto (BD Bioscience) and analysed by using the FLOWJO software (Tree Star).

BMDM and BMDC cultures. For BMDM differentiation, primary bone marrow cells were cultured for 6 days in DMEM supplemented with 50 U ml^{-1} penicillin, 50 μg ml^{-1} streptomycin (all from Invitrogen), 30% (vol/vol) supernatant from L929 cells and 10% (vol/vol) fetal calf serum (PAA). BMDCs were generated by incubating primary bone marrow cells for 7–8 days in RPMI 1640 with 1% (vol/vol) nonessential amino acids, 1% (vol/vol) sodium pyruvate, 50 U ml^{-1} penicillin, 50 μg ml^{-1} streptomycin, 5×10^{-5} M 2-mercaptoethanol (all from Invitrogen) and 10% (vol/vol) fetal calf serum (PAA), supplemented with 20 ng ml^{-1} recombinant GM-CSF (from ImmunoTools).

***In vitro* generation/restimulation of effector CD4$^+$ and CD8$^+$ T cells.** For effector T-cell generation, sorted naive CD4$^+$ or CD8$^+$ cells were stimulated *in vitro* in RPMI 1640 with 1% (vol/vol) nonessential amino acids, 1% (vol/vol) sodium pyruvate, 50 U ml^{-1} penicillin, 50 μg ml^{-1} streptomycin, 5×10^{-5} M 2-mercaptoethanol (all from Invitrogen) and 10% (vol/vol) fetal calf serum (from PAA), with 10 μg ml^{-1} plastic-bound functional grade purified anti-CD3 (145-2C11), 2 μg ml^{-1} soluble anti-CD28 (37.51) (both antibodies were from eBioscience) and 5 ng ml^{-1} human recombinant IL-2 (Glaxo). Effector cells were collected on day 5, or as indicated.

For restimulation of effector T cells, effector cells were restimulated overnight *in vitro* with 10 μg ml^{-1} plastic-bound functional grade purified anti-CD3, 2 μg ml^{-1} soluble anti-CD28 and 5 ng ml^{-1} human recombinant IL-2. Fixation with glutaraldehyde (from Sigma) was performed as described elsewhere[19].

***In vitro* stimulation experiments.** The medium used was RPMI 1640 with 1% (vol/vol) nonessential amino acids, 1% (vol/vol) sodium pyruvate, 50 U ml^{-1} penicillin, 50 μg ml^{-1} streptomycin, 5×10^{-5} M 2-mercaptoethanol (all from Invitrogen) and 10% (vol/vol) fetal calf serum (from PAA). Recombinant mouse CD200–Fc and mouse IFN-γ were purchased from R&D Systems.

ELISA. Cell culture supernatants were assayed for mouse IL-1β, CXCL2, TNF (R&D Systems), IL-18 (MBL International), IL-6 and IL-12 (eBioscience) according to the manufacturer's instructions.

Immunoblot analysis. The antibody against mouse IL-1β was a gift from R. Solari. The antibody against mouse caspase-1 was a gift from P. Vandenabeele. Rabbit polyclonal antibody to β-actin was purchased from Abcam.

***In vivo* immunization experiments.** Mice were immunized by i.p. injection of 50 μg CpG oligodeoxynucleotide 1826 (5′-CCATGACGTTCCTGACGTT-3′; Microsynth), together with 100 μg OVA, from Calbiochem. Ten to eleven days after immunization, mice were injected i.p. either first with 100 μg $OVA_{protein}$ followed 4 h later by a second injection of 250 μg alum together with 600 μg $OVA_{protein}$, or with 250 μg alum alone. Between 12 and 14 h after alum injection, mice were killed and peritoneal cavities washed with PBS. Neutrophil (CD11b$^+$ Ly-6C$^+$ Ly-6G^{high} F4/80$^-$) and eosinophil (CD11b$^+$ Ly-6C^{int} Ly-6G^{int} F4/80^{int}) recruitment was analysed by FACS.

Statistical analysis. For *in vitro* experiments, statistical analyses were calculated with a paired Student's *t*-test; for *in vivo* experiments, differences between groups were calculated using the Mann–Whitney U test for unpaired data (GraphPad Prism version 5.0). Differences were considered significant when $P \leq 0.05$ (asterisk), very significant when $P \leq 0.01$ (two asterisks) and extremely significant when $P \leq 0.001$ (three asterisks).

26. Wang, Y. *et al.* Th2 lymphoproliferative disorder of LatY136F mutant mice unfolds independently of TCR–MHC engagement and is insensitive to the action of Foxp3$^+$ regulatory T cells. *J. Immunol.* **180**, 1565–1575 (2008).
27. Kawabe, T. *et al.* The immune responses in CD40-deficient mice: impaired immunoglobulin class switching and germinal center formation. *Immunity* **1**, 167–178 (1994).
28. Kuida, K. *et al.* Altered cytokine export and apoptosis in mice deficient in interleukin-1 beta converting enzyme. *Science* **267**, 2000–2003 (1995).
29. Chen, J. *et al.* Immunoglobulin gene rearrangement in B cell deficient mice generated by targeted deletion of the JH locus. *Int. Immunol.* **5**, 647–656 (1993).

Vol 460|9 July 2009|doi:10.1038/nature08102

LETTERS

Unlimited multistability in multisite phosphorylation systems

Matthew Thomson[1] & Jeremy Gunawardena[2]

Reversible phosphorylation on serine, threonine and tyrosine is the most widely studied posttranslational modification of proteins[1,2]. The number of phosphorylated sites on a protein (n) shows a significant increase from prokaryotes, with $n \leq 7$ sites, to eukaryotes, with examples having $n \geq 150$ sites[3]. Multisite phosphorylation has many roles[4,5] and site conservation indicates that increasing numbers of sites cannot be due merely to promiscuous phosphorylation. A substrate with n sites has an exponential number (2^n) of phospho-forms and individual phospho-forms may have distinct biological effects[6,7]. The distribution of these phospho-forms and how this distribution is regulated have remained unknown. Here we show that, when kinase and phosphatase act in opposition on a multisite substrate, the system can exhibit distinct stable phospho-form distributions at steady state and that the maximum number of such distributions increases with n. Whereas some stable distributions are focused on a single phospho-form, others are more diffuse, giving the phosphoproteome the potential to behave as a fluid regulatory network able to encode information and flexibly respond to varying demands. Such plasticity may underlie complex information processing in eukaryotic cells[8] and suggests a functional advantage in having many sites. Our results follow from the unusual geometry of the steady-state phospho-form concentrations, which we show to constitute a rational algebraic curve, irrespective of n. We thereby reduce the complexity of calculating steady states from simulating 3×2^n differential equations to solving two algebraic equations, while treating parameters symbolically. We anticipate that these methods can be extended to systems with multiple substrates and multiple enzymes catalysing different modifications, as found in posttranslational modification 'codes'[9] such as the histone code[10,11]. Whereas simulations struggle with exponentially increasing molecular complexity, mathematical methods of the kind developed here can provide a new language in which to articulate the principles of cellular information processing[12].

A major difficulty in studying multisite phosphorylation from a systems perspective has been the lack of information regarding when different sites on the same protein are simultaneously phosphorylated[13]. Sites may be intricately dependent on each other[14] and both the number and the position of phosphates can affect biological outcome[6,7]. The phospho-form distribution—the relative stoichiometries of each of the 2^n phospho-forms—is thus the appropriate measure of phosphorylation state for a multisite substrate. Such distributions are starting to be measured[15,16], prompted by interest in posttranslational modification (PTM) codes. A theoretical understanding will improve our ability to interpret such data, uncover biological principles and design appropriate experiments. Although we focus on phosphorylation, we hope to lay a foundation for analysing multiple PTMs.

Figure 1 summarizes a general model of multisite phosphorylation. A substrate S, with n sites, is acted on by a kinase E and a phosphatase F. Substrates may have multiple kinases and phosphatases *in vivo* but a single enzyme often addresses multiple sites and we focus here on the minimal enzymatic machinery needed for any n. Each enzyme may act distributively (Fig. 1a) as well as processively (Fig. 1b) using a standard biochemical scheme (Fig. 1c) and arbitrary preferences for site order. These assumptions are more general than in previous models[17–22]: by choosing the details appropriately, any kinase, phosphatase and substrate system can be represented. ATP is assumed to be recharged by some process that is not modelled. It is therefore not treated as a variable but its effect is absorbed into the site-specific parameters (a_u^X, b_u^X, $c_{u,v}^X$ defined in Fig. 1c). Phosphorylation and dephosphorylation are assumed to take place on a fast time scale in comparison to synthesis and degradation of the component proteins. The model is, therefore, effectively closed: there is no flux of material through it and the total amounts of substrate, S_{tot}, and enzymes, E_{tot} and F_{tot}, remain constant at all times. With mass-action kinetics, these assumptions give rise to 3×2^n nonlinear differential equations for the state variables (Fig. 1d).

With limited information on site-specific parameters, numerical simulations can be undertaken for randomly selected parameter values in an attempt to discern typical behaviours. However, doing so in a state space of dimension 3×2^n rapidly becomes intractable as n increases. Here we introduce a new method of analysis, which allows strong conclusions to be drawn about steady states without having to specify parameter values in advance. Experimental evidence indicates that biological systems attain quasi-steady states *in vivo*[23,24], including systems in which multisite phosphorylation has a significant role[25]. While the steady-state assumption must be confirmed in each experimental context, it has been widely used in modelling multisite phosphorylation[17–22].

At steady state, the second equation in Fig. 1d yields an expression for XS_u in terms of X and S_u. Substituting this into the first equation in Fig. 1d makes that equation linear in S_u. The coefficients are algebraic expressions in the site-specific parameters (collectively denoted $\mathbf{a}$) and the auxiliary parameter $t = E/F$, the steady-state ratio of free kinase to free phosphatase. These expressions may be regarded as elements of a set of coefficients, $\mathbb{R}(\mathbf{a}, t)$, in which the $\mathbf{a}$ values and the t have been adjoined to the ordinary numbers, $\mathbb{R}$, as uninterpreted symbols that can be added, subtracted, multiplied and divided as if they were numbers (see the Supplementary Information for more details). The elements of $\mathbb{R}(\mathbf{a}, t)$ correspond to rational functions, or ratios of polynomials, in these parameters (Supplementary Information). By treating the parameters symbolically in this way, they can be used in calculations without their numerical values being known in advance.

The linearized equations for the S_u can be solved by Gaussian elimination, which works as well over the coefficients $\mathbb{R}(\mathbf{a}, t)$ as over $\mathbb{R}$ (Supplementary Information). The steady-state phospho-forms can thereby be shown to satisfy (Supplementary Information)

[1]Biophysics Program, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.

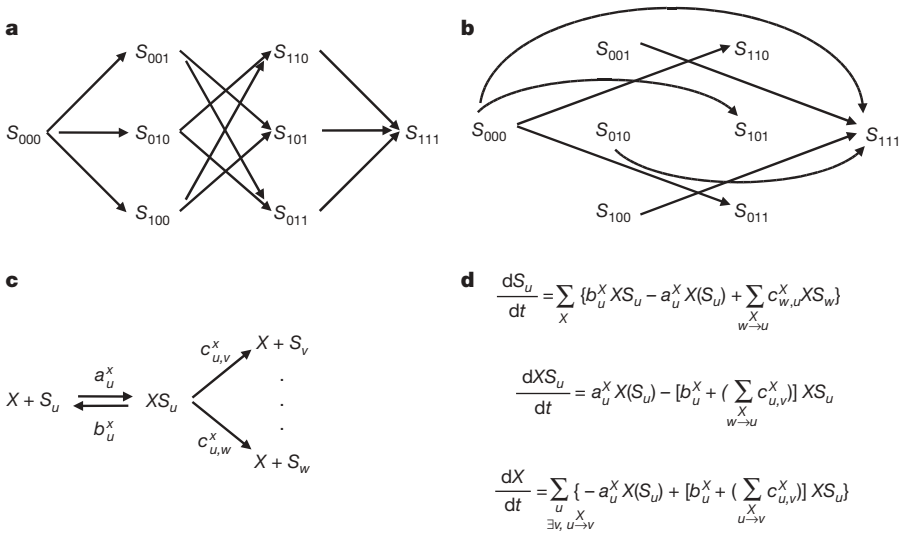


Figure 1 | General model of multisite phosphorylation with substrate *S* having *n* sites, kinase *E* and phosphatase *F*. The phospho-forms are denoted S_u, where u is a bit string indicating the absence/presence (0/1) of phosphate. Kinase reactions for $n = 3$ with one phosphorylation per reaction (**a**, distributivity[29]) and multiple phosphorylations per reaction (**b**, processivity[29]). Phosphatases act similarly in removing phosphates. **c**, Each enzyme ($X = E$ or F) uses a standard biochemical mechanism[30] but may form multiple products, with associated parameters (a_u^X, b_u^X, $c_{u,v}^X$) for mass-action kinetics. ATP is assumed to be held constant and synthesis and degradation are ignored. **d**, The 3×2^n differential equations, where $u \xrightarrow{X} v$ signifies that X converts S_u to S_v. The same symbol is used for a chemical species and for its concentration. Note that $X(S_u)$ here indicates the product of X and S_u.

$$S_u = S_0 r_u(t) \quad (1)$$

where S_0 is the unphosphorylated phospho-form and $r_u(t)$ is a rational function of t with coefficients in $\mathbb{R}(\mathbf{a})$. Although these rational functions are complex, they can be explicitly calculated for any given model (Supplementary Information).

If numerical values are to be given to the site-specific parameters, it is important to know that the rational functions $r_u(t)$ remain well defined. For instance, the rational function $c/(1-c)$ becomes undefined when $c = 1$. We show that, for any positive site-specific parameter values and positive t, $r_u(t)$ is always well defined and positive ('positivity', Supplementary Information), ruling out such problems.

Equation (1) implies that the steady-state phospho-forms can be described through a single auxiliary variable, t, so that they form a one-dimensional geometric object, or 'curve'. Despite increasing numbers of sites, and the exponentially increasing size of the model, the steady-state phospho-forms always remain a curve, providing the basis for an exponential reduction in complexity. What changes with n is the extent to which the curve undulates, which determines how many steady states can co-exist for given amounts of substrate and enzymes (see below). Not all curves can be described by rational functions; those that can are of considerable geometric interest, as explored in an earlier paper[26] and discussed further in the Supplementary Information.

The upshot of equation (1) is that, at steady state, the 3×2^n state variables are determined by S_0, E and F. Because the substrate and enzyme totals remain constant, we can formulate three equations for these three unknowns. The equation for S_0 can be solved directly in terms of S_{tot}, leaving a pair of equations (defined in the Supplementary Information):

$$\Phi_1(E, F) = E_{\mathrm{tot}} \quad \text{and} \quad \Phi_2(E, F) = F_{\mathrm{tot}} \quad (2)$$

which determine the steady-state E and F values corresponding to any given substrate and enzyme totals. Equation (2) exactly characterizes the steady states of the model (Supplementary Information). To find steady states, it is no longer necessary to numerically simulate 3×2^n differential equations; this can be done by only solving two algebraic equations. The complexity arising from the dynamics has been distilled away. This exponential reduction of complexity is the key to what follows.

Figure 2a shows an example with four sites and five steady states. (We assume sequentiality, as in Fig. 3a, but merely for convenience.) Only stable states are detected experimentally, or found by numerical simulation, and this example is tristable; the corresponding stable phospho-form distributions are markedly distinct. Whereas distributions 1 and 3 are each focused on a single phospho-form, distribution 2 is broader. With multiple stable states a system can encode many outcomes, or several bits of information, enabling complex information encoding and processing[8]. Such multistability is believed to underlie cellular differentiation and other decisions[8,27] but experimental examples have, so far, only demonstrated bistability[23–25]. Bistability in two-site phosphorylation was previously shown by modelling[20].

It remained unknown how the number of stable states depends on n. We found that multistability tends to occur when substrate is in excess over enzymes (as in Fig. 2a). In this approximation, the two equations for E and F in equation (2) can be reduced to one polynomial equation for $t = E/F$ (Supplementary Information):

$$P(t) = \alpha_N t^N + \alpha_{N-1} t^{N-1} + \cdots + \alpha_1 t + \alpha_0 = 0 \quad (3)$$

where $\alpha_i \in \mathbb{R}(\mathbf{a})$ and N lies between $n + 1$ and 2^n depending on the model. Positivity of the $r_u(t)$ is essential here. Positive solutions of $P(t) = 0$ correspond approximately to steady states. We show that the discrepancy between the exact steady states found by equation (2) and the approximate ones found by equation (3) can be made as small as desired by increasing the excess of S_{tot} over E_{tot} and F_{tot} (Supplementary Information).

The advantage of equation (3) over equation (2) is that we can readily construct solutions of the former. If n is even and we choose any $n + 1$ distinct positive numbers, then we can always find a model for which the corresponding $P(t)$ has these numbers as solutions (Supplementary Information). Moreover, S_{tot} can be chosen arbitrarily, so that the approximation of equation (2) by equation (3) is as accurate as desired. It follows that the corresponding model has $n + 1$ steady states. The example in Fig. 2a meets this bound for $n = 4$. If n is odd the same can be achieved for any set of n distinct positive numbers (Supplementary Information). Note that odd numbers of steady states are constructed in both cases.

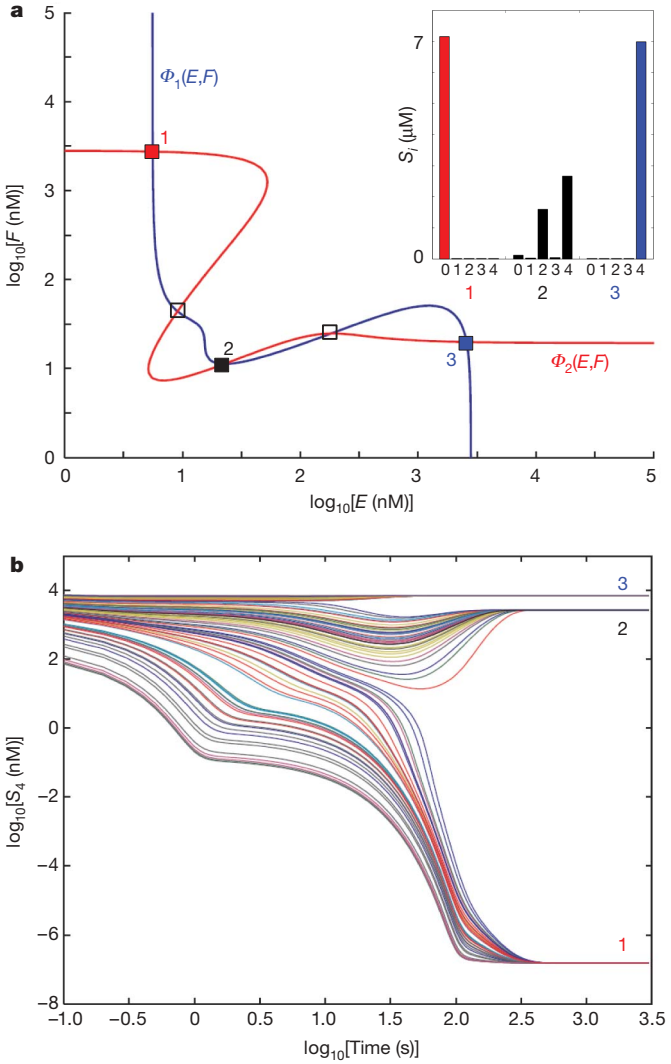


Figure 2 | Multistability for an $n = 4$ distributive, sequential system, as in Fig. 3a. For parameter values, see Supplementary Table 1. **a**, Plots of equation (2) for $S_{tot} = 10\,\mu M$ and $E_{tot} = F_{tot} = 2.8\,\mu M$. The intersections correspond to the steady states. Filled squares are stable: 1 (red), 2 (black) and 3 (blue); open squares are unstable. Stability was determined by standard methods (Supplementary Information). The inset shows the corresponding phospho-form distributions, following the notation in Fig. 3a. **b**, Time courses of S_4 reaching its three stable values from initial conditions $S_0 = \alpha S_{tot}$, $S_4 = (1-\alpha)S_{tot}$ and $X = 0$ for all other variables, with α chosen randomly from the uniform distribution on [0, 1] (100 samples), determined by simulation (Supplementary Information).

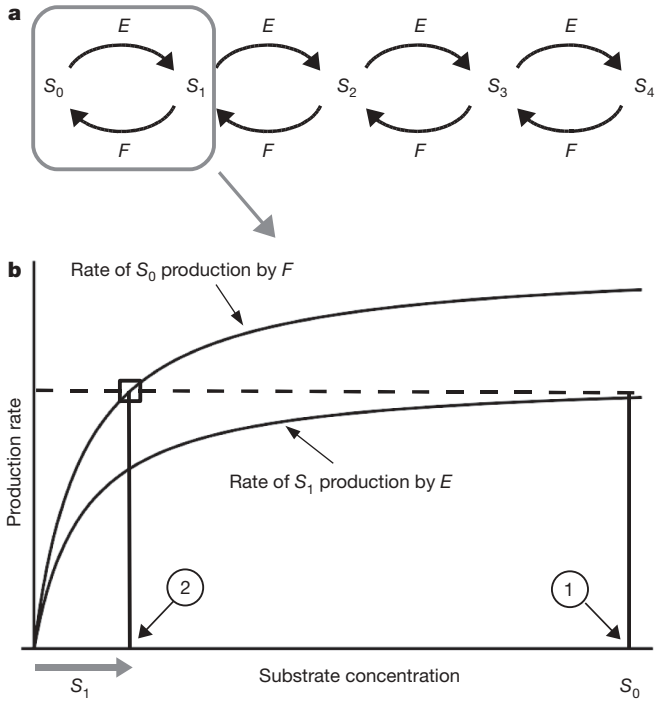


Figure 3 | Multistability by kinetic trapping. **a**, A distributive, sequential system with $n = 4$; E phosphorylates in order, F dephosphorylates in reverse order. The phospho-forms are denoted S_i, $i = 0,\ldots, 4$, where i is the number of phosphates. **b**, Rate functions for production of S_1 from S_0 by E and of S_0 from S_1 by F are approximated as Michaelis–Menten hyperbolas[30], with the phosphatase curve to the left and above the kinase curve. If the system is initiated with substrate entirely in S_0 and in excess over both enzymes and saturating them, S_0 will sequester E and produce S_1 at nearly maximal rate (point 1). F, however, will be unoccupied, so that as S_1 increases (grey arrow), phosphorylation and dephosphorylation will balance (point 2). Any leak of S_1 into S_2 can be opposed by F, which is not sequestered. The system will hence reach steady state with substrate remaining predominantly unphosphorylated. If similar conditions are applied to S_3 and S_4, but reversed with respect to E and F, then S_4 can be similarly trapped. The parameter values required for this argument are discussed in the Supplementary Information.

This is also what we find in simulations. Furthermore, if the steady states are ordered by their corresponding E/F values then unstable states always occur between stable ones (Supplementary Information), as in Fig. 2a. Hence, if there are $2k + 1$ steady states, $k + 1$ of them are stable. It follows that a model with n sites can have as many as $\lfloor (n+2)/2 \rfloor$ stable states ($\lfloor x \rfloor$ being the greatest integer not greater than x). We see that the tristability in Fig. 2a is only the tip of the iceberg: the maximum number of stable states increases with increasing numbers of sites.

Experimental detection of stable states requires an understanding of how they arise dynamically. Figure 3b considers the inter-conversion of S_0 and S_1, for the sequential system in Fig. 3a. An informal argument based on the Michaelis–Menten approximation suggests that, for suitable parameter values, if the system is started entirely in S_0 then substrate can remain trapped predominantly in that state, as in distribution 1 in Fig. 2a. A similar argument applies to trapping of S_4, as in distribution 3. Simulations show that all three stable distributions can be reached by starting from suitable mixtures of S_0 and S_4 (Fig. 2b).

In vivo, the most likely way that a phosphorylation system is regulated is by modulating its enzymes. Because the enzymes have been assumed to be in their active states, this corresponds to altering E_{tot} or F_{tot}. Figure 4a shows, for the tristable system in Fig. 2, that changes in E_{tot} can switch the system between stable states 1 and 3. It is possible that more complex modulations, involving both E_{tot} and F_{tot}, could access all three stable states or that different parameter values could facilitate additional switching capability. Figure 4b shows another option, in which changes in E_{tot} can switch between three stable states even though there is only a narrow window of E_{tot} values for which three stable states coexist. In other words, the system may not need to exhibit robust tristability to have access to three stable states.

DNA provides a static, structural mechanism for encoding information at a capacity of 2 bits per base pair. The idea that PTMs provide a dynamic mechanism for information encoding has been broadly influential[10,11] but the mechanistic details remain a matter of debate[9,28] and no estimate of information capacities has emerged for any such code. Our result provides the first demonstration of a PTM mechanism that can, in principle, encode an arbitrary amount of information, along with an estimate of its information capacity. If natural selection has found such a capability useful, that may help account for the emergence of large numbers of phosphorylation sites.

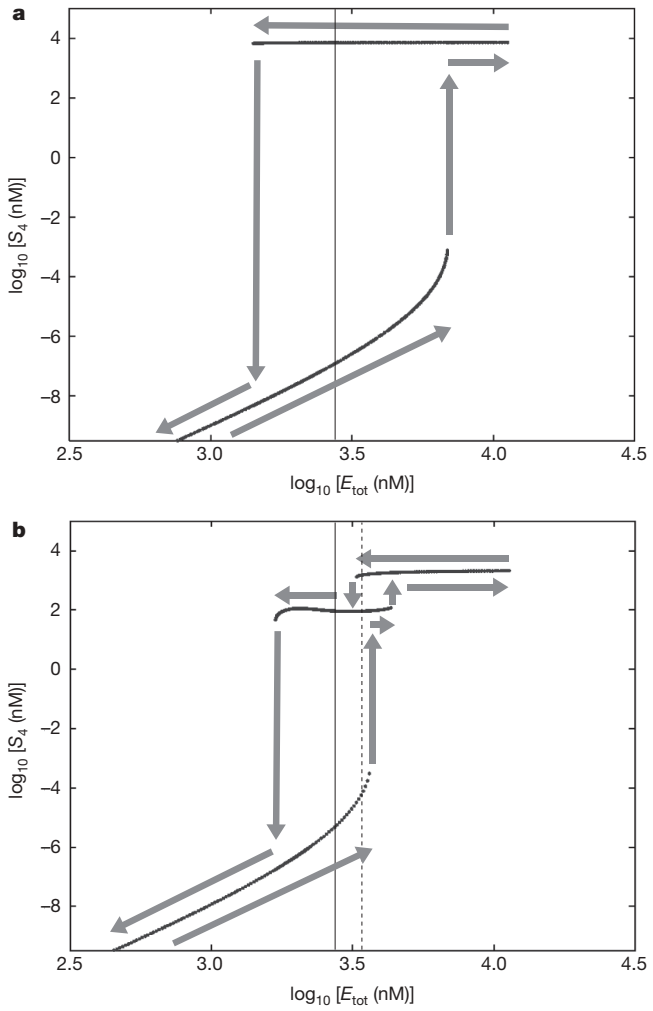


Figure 4 | Switching between stable states. **a**, The system in Fig. 2a with $E_{tot} = F_{tot} = 2.8\ \mu M$ and $S_{tot} = 10\ \mu M$ is taken in a cycle (grey arrows) using simulation (Supplementary Information). The free kinase is repeatedly changed by a small amount and the system allowed to relax back to a stable state. Starting on the lower branch, corresponding to distribution 1 in Fig. 2a, the system switches abruptly to the higher branch (distribution 3), remains on that branch as E_{tot} is lowered, and then switches abruptly back down to the lower branch (hysteresis). **b**, The system with $E_{tot} = F_{tot} = 2.8\ \mu M$, $S_{tot} = 5\ \mu M$ has only two stable states (not shown) but occupies three when E_{tot} is cycled. Three stable states only coexist in a narrow window around the dotted line. The solid lines mark $E_{tot} = 2.8\ \mu M$.

Received 10 November 2008; accepted 24 April 2009.
Published online 17 June 2009.

1. Walsh, C. T. *Posttranslational Modification of Proteins* (Roberts and Company, 2006).
2. Cohen, P. The role of reversible protein phosphorylation in health and disease. *Eur. J. Biochem.* **268,** 5001–5010 (2001).
3. Gnad, F. *et al.* PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* **8,** R250 (2007).
4. Cohen, P. The regulation of protein function by multisite phosphorylation — a 25 year update. *Trends Biochem. Sci.* **25,** 596–601 (2000).
5. Holmberg, C. I., Tran, S. E. F., Eriksson, J. E. & Sistonen, L. Multisite phosphorylation provides sophisticated regulation of transcription factors. *Trends Biochem. Sci.* **27,** 619–627 (2002).
6. Wu, R. C. *et al.* Selective phosphorylations of the SRC-3/AIB1 coactivator integrate genomic responses to multiple cellular signaling pathways. *Mol. Cell* **15,** 937–949 (2004).
7. Park, K.-S., Mohapatra, D. P., Misonou, H. & Trimmer, J. S. Graded regulation of the Kv2.1 potassium channel by variable phosphorylation. *Science* **313,** 976–979 (2006).
8. Nurse, P. Life, logic and information. *Nature* **454,** 424–426 (2008).
9. Sims, R. J. & Reinberg, D. Is there a code embedded in proteins that is based on posttranslational modification? *Nature Rev. Mol. Cell Biol.* **9,** 815–820 (2008).
10. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293,** 1074–1080 (2001).
11. Turner, B. Cellular memory and the histone code. *Cell* **111,** 285–291 (2002).
12. Cohen, J. E. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol.* **2,** e439 (2004).
13. Hunter, T. The age of crosstalk: phosphorylation, ubiquitination and beyond. *Mol. Cell* **28,** 730–738 (2007).
14. Ferrarese, A. *et al.* Chemical dissection of the APC repeat 3 multistep phosphorylation by the concerted action of protein kinases CK1 and GSK3. *Biochemistry* **46,** 11902–11910 (2007).
15. Phanstiel, D. *et al.* Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells. *Proc. Natl Acad. Sci. USA* **105,** 4093–4098 (2008).
16. Pesavento, J. J., Bullock, C. R., LeDuc, R. D., Mizzen, C. A. & Kelleher, N. L. Combinatorial modification of human histone H4 quantitated by two-dimensional liquid chromatography coupled with top down mass spectrometry. *J. Biol. Chem.* **283,** 14927–14937 (2008).
17. Goldbeter, A. & Koshland, D. E. An amplified sensitivity arising from covalent modification in biological systems. *Proc. Natl Acad. Sci. USA* **78,** 6840–6844 (1981).
18. Lisman, J. E. A mechanism for memory storage insensitive to molecular turnover: a bistable autophosphorylating kinase. *Proc. Natl Acad. Sci. USA* **82,** 3055–3057 (1985).
19. Salazar, C. & Höfer, T. Allosteric regulation of the transcription factor NFAT1 by multiple phosphorylation sites: a mathematical analysis. *J. Mol. Biol.* **327,** 31–45 (2003).
20. Markevich, N. I., Hoek, J. B. & Kholodenko, B. N. Signalling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J. Cell Biol.* **164,** 353–359 (2004).
21. Gunawardena, J. Multisite protein phosphorylation makes a good threshold but can be a poor switch. *Proc. Natl Acad. Sci. USA* **102,** 14617–14622 (2005).
22. Kim, S. Y. & Ferrell, J. E. Substrate competition as a source of ultrasensitivity in the inactivation of Wee1. *Cell* **128,** 1133–1145 (2007).
23. Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I. & van Oudenaarden, A. Multistability in the lactose utilization network of *Escherichia coli*. *Nature* **427,** 737–740 (2004).
24. Sha, W. *et al.* Hysteresis drives cell-cycle transitions in *Xenopus laevis* egg extracts. *Proc. Natl Acad. Sci. USA* **100,** 975–980 (2003).
25. Ferrell, J. E. Jr & Machleder, E. M. The biochemical basis of an all-or-none cell fate switch in *Xenopus* oocytes. *Science* **280,** 895–898 (1998).
26. Manrai, A. & Gunawardena, J. The geometry of multisite phosphorylation. *Biophys. J.* **95,** 5533–5543 (2008).
27. Monod, J. & Jacob, F. General conclusions: teleonomic mechanisms in cellular metabolism, growth and differentiation. *Cold Spring Harb. Symp. Quant. Biol.* **26,** 389–401 (1961).
28. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447,** 407–411 (2007).
29. Huang, C.-Y. F. & Ferrell, J. E. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc. Natl Acad. Sci. USA* **93,** 10078–10083 (1996).
30. Cornish-Bowden, A. *Fundamentals of Enzyme Kinetics* 2nd edn 23–28 (Portland Press, 1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported in part by the NIH under grant R01-GM081578. We thank A. Manrai for scientific discussions, R. Ward for editorial help and HMS RITG for support with cluster computing. We acknowledge the encouragement of the late Stephen Thomson (1946–2006) and Charles Gunawardena (1929–2007).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.G. (jeremy@hms.harvard.edu).

Vol 460|9 July 2009|doi:10.1038/nature08136

LETTERS

A mechanism linking extra centrosomes to chromosomal instability

Neil J. Ganem[1], Susana A. Godinho[1] & David Pellman[1]

Chromosomal instability (CIN) is a hallmark of many tumours and correlates with the presence of extra centrosomes[1–4]. However, a direct mechanistic link between extra centrosomes and CIN has not been established. It has been proposed that extra centrosomes generate CIN by promoting multipolar anaphase, a highly abnormal division that produces three or more aneuploid daughter cells. Here we use long-term live-cell imaging to demonstrate that cells with multiple centrosomes rarely undergo multipolar cell divisions, and the progeny of these divisions are typically inviable. Thus, multipolar divisions cannot explain observed rates of CIN. In contrast, we observe that CIN cells with extra centrosomes routinely undergo bipolar cell divisions, but display a significantly increased frequency of lagging chromosomes during anaphase. To define the mechanism underlying this mitotic defect, we generated cells that differ only in their centrosome number. We demonstrate that extra centrosomes alone are sufficient to promote chromosome missegregation during bipolar cell division. These segregation errors are a consequence of cells passing through a transient 'multipolar spindle intermediate' in which merotelic kinetochore–microtubule attachment errors accumulate before centrosome clustering and anaphase. These findings provide a direct mechanistic link between extra centrosomes and CIN, two common characteristics of solid tumours. We propose that this mechanism may be a common underlying cause of CIN in human cancer.

A conspicuous feature of many tumour cells is an increased rate of gain or loss of whole chromosomes, a phenomenon referred to as chromosomal instability (CIN)[1]. Cells with CIN missegregate chromosomes 10–100 times more frequently than non-transformed or chromosomally stable diploid cancers cells. CIN is thus a major source of aneuploidy[1,5] and has important implications not only for tumour initiation, where aneuploidy can have a causal role[6], but also for tumour cell evolution, where increased rates of chromosome missegregation may enable clonal expansion of cells with proliferative advantages, metastatic potential or chemoresistance[5,7,8]. Despite its importance, the mechanisms leading to CIN in most cancers are not defined.

One proposed mechanism underlying CIN is extra-centrosome-mediated multipolar spindle assembly followed by asymmetric chromosome segregation resulting in massive aneuploidy[3]. Correlative support for this idea comes from the fact that extra centrosomes and multipolar mitotic figures are common in CIN cancers, yet rare in chromosomally stable tumours[9–12]. However, an obvious paradox arises when considering such multipolar cell division as an underlying mechanism of CIN: aneuploidy compromises cell fitness[6,13,14] and therefore massive aneuploidy following multipolar cell division would probably compromise viability[3,15–17]. Until now, neither the frequency of multipolar divisions nor the fate of the resulting progeny has been systematically characterized in cancer cells.

To visualize directly the relationship between multipolar cell division and cell viability, we generated a variety of cancer cell lines from different tissues of origin that stably express the chromosome marker histone H2B–GFP (green fluorescent protein) and performed long-term live-cell imaging. The percentage of cells that harboured extra centrosomes varied significantly, from ~12% in HT-29 human colon cancer cells to 100% in mouse neuroblastoma N1E-115 cells (Fig. 1a). Nevertheless, our analysis clearly indicated that the fraction of cells undergoing multipolar cell division, defined by the segregation of chromosomes to three or more poles during anaphase, was always markedly less than the fraction of cells possessing extra centrosomes (Fig. 1a). Thus, as expected from previous work, clustering of supernumerary centrosomes into two poles is an efficient mechanism that limits multipolar cell division[3,15–19].

We determined the fate of cells that underwent a multipolar anaphase. We observed that, although multipolar anaphase often produced three or more mononucleated daughter cells, cytokinesis failure along one or both division planes to produce binucleate or polynucleate progeny was also common (Supplementary Fig. 1a). Both mono- and poly-nucleated progeny (P1) of spontaneously arising multipolar divisions were then tracked over a four-day period and classified as: undergoing mitotic cell death with a multipolar spindle configuration; dying during the subsequent interphase; undergoing cell-cycle arrest; or successfully completing a second round of cell division to generate P2 progeny (Fig. 1b, c). Strikingly, we found that most progeny (P1) of multipolar cells died or arrested, regardless of tissue of origin or whether the cells were mono- or poly-nucleated (Fig. 1b–d, Supplementary Fig. 1a, b and Supplementary Movie 1). Moreover, when rare P1 progeny from multipolar divisions completed a second round of mitosis to generate P2 progeny, even fewer of the resulting daughter cells were capable of further division (Fig. 1b, c). Finally, we observed that progeny from multipolar divisions were usually inviable even if they were born from binucleated, presumably tetraploid, mother cells. This indicates that doubling the chromosome content does not efficiently buffer the deleterious effects of massive aneuploidy that results from multipolar anaphase (Supplementary Fig. 2).

Taken together, these data provide two reasons why multipolar cell division alone cannot explain the high rates of chromosome missegregation in CIN cells. First, the frequencies of multipolar division are not high enough to account for the observed rates of chromosome missegregation in these cell lines; for example, whereas MCF-7 and HT-29 cells missegregate chromosomes on average once every 2 or 5 divisions, respectively[1,13], they each undergo multipolar anaphase only once every 50 or more divisions. Second, the failure of progeny to continue proliferating reveals that multipolar mitosis cannot give rise to persistently unstable cells.

We therefore considered other mechanisms by which multiple centrosomes might generate CIN. Recent work has revealed that

[1]Howard Hughes Medical Institute, Department of Pediatric Oncology, Dana-Farber Cancer Institute, Children's Hospital, Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.

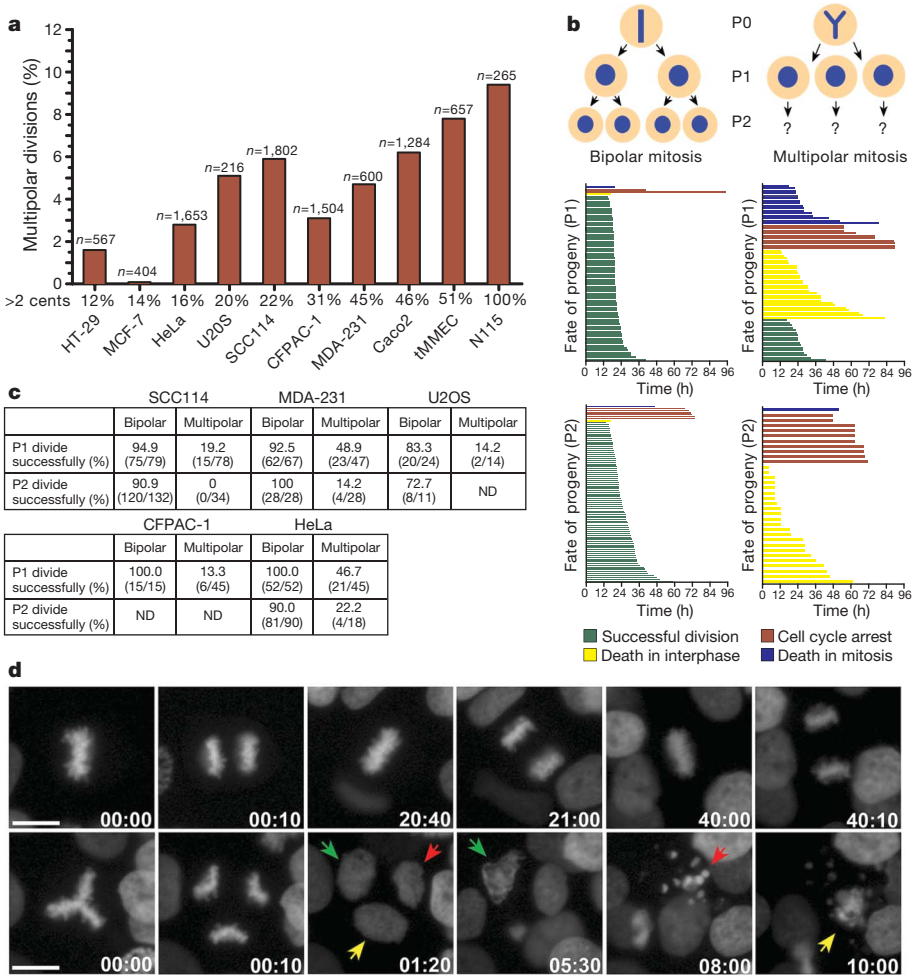


Figure 1 | Multipolar cell divisions are rare and the progeny are typically inviable. **a**, The percentage of different cancer cells that undergo multipolar cell division based on live-cell imaging (n = number of cell divisions). Cents, centrosomes. **b**, A representative cell fate analysis from SCC114 cells. Individual cell fates of progeny from both bipolar (left column) and multipolar (right column) cell divisions (represented by single coloured lines) are shown. **c**, Percentage of progeny from bipolar and multipolar cell divisions that undergo successful cell division. ND, not determined. **d**, Still frames from the imaging experiment represented in **b** showing a representative SCC114 cell undergoing several rounds of bipolar cell division (top row) or a single multipolar cell division (bottom row). Coloured arrows track the fate of the three progeny from the multipolar cell division (Supplementary Movie 1). Time, hours:minutes. Scale bars, 10 μm.

merotely, a type of error in which single kinetochores attach to microtubules emanating from different poles[20,21], is common in CIN cells[13]. Merotelic attachments are particularly dangerous because they are poorly sensed by the spindle assembly checkpoint and, if not corrected, may give rise to lagging chromosomes during anaphase that can lead to missegregation events[20–23]. However, the cause of merotely in CIN cancers is unknown.

We hypothesized that cells with extra centrosomes pass through transient multipolar intermediates before centrosome clustering and that the geometry of such intermediates predisposes to merotelic attachments. Supporting this idea, we found that CIN cells with extra centrosomes spend most of mitosis in a multipolar configuration before centrosome clustering (Supplementary Fig. 3), and, furthermore, we directly observed numerous merotelic attachments within multipolar spindles by high-resolution microscopy (Fig. 2a and Supplementary Fig. 8). Moreover, as observed in tetraploid PtK_1 cells[24,25], both the frequency and number of lagging chromosomes are notably increased in rare cells that undergo multipolar anaphase, indicating that merotelic attachments are enriched within the multipolar configuration (Supplementary Fig. 4 and Supplementary Movie 2).

As a further test of the hypothesis, a panel of CIN cell lines was analysed by fixed-cell imaging and scored for both lagging chromosomes (markers for merotelic attachment) and centrosome number during mid-anaphase (Fig. 2b, c). The presence of extra centrosomes during bipolar anaphase correlated with a significant increase (threefold to tenfold) in the frequency of lagging chromosomes in every cell line examined, consistent with the idea that extra centrosomes increase the formation of merotelic attachments that can persist through anaphase (Fig. 2b, c).

We next tested whether extra centrosomes alone are sufficient to promote chromosome missegregation during bipolar cell division. A significant technical challenge for studying the consequences of centrosome amplification has been the difficulty in obtaining genetically matched cells that do or do not contain extra centrosomes. We circumvented this problem in two ways. First, we generated tetraploid cells with extra centrosomes by using cytochalasin D to inhibit cytokinesis in chromosomally stable non-transformed telomerase-immortalized human BJ and RPE-1 cell lines. Because non-transformed cells are prone to cell-cycle arrest after tetraploidization, we transiently knocked down p53 (also known as TP53) by small interfering RNA (siRNA) to facilitate passage of tetraploids through mitosis. Like cancer cells, tetraploid cells clustered their extra centrosomes into two poles during the subsequent mitosis and displayed a significantly increased frequency of anaphase lagging chromosomes during bipolar anaphase relative to matched diploids that were also exposed to cytochalasin and depleted of p53 (Fig. 3a, c).

Interestingly, both tetraploid BJ and RPE-1 cells spontaneously lost their extra centrosomes after passage in culture (Fig. 3a and data

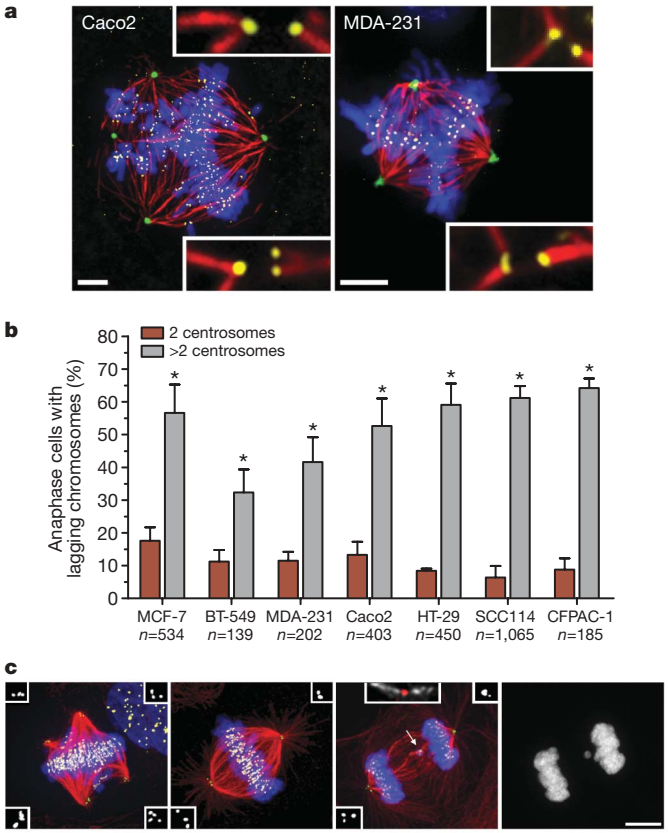


Figure 2 | Extra centrosomes correlate with increases in lagging chromosomes. **a**, Multipolar Caco2 and MDA-231 cells stained for pericentrin (green), microtubules (red), chromosomes (blue) and kinetochores (yellow). Merotelic attachments are shown in insets. Scale bars, 5 μm. **b**, The percentage of cancer cells with two (red) or more than two (grey) centrosomes that exhibit one or more lagging chromosomes during bipolar anaphase (*n*, number of anaphase cells counted); error bars represent mean ± s.e.m. from at least 4 independent experiments; asterisks denote *P* values <0.02 and are derived from an unpaired two-tailed *t*-test. **c**, Representative MCF-7 cells with extra centrosomes during prometaphase (multipolar spindle intermediate, left panel), metaphase (bipolar spindle with clustered centrosomes, second panel from left) and anaphase (with clustered centrosomes, third and fourth panels from left), stained for centrioles (green, and inset white), microtubules (red), chromosomes (blue) and centromeres (yellow). Arrow indicates a lagging chromosome caused by merotelic attachment, in which microtubules emanating from both poles attach to a single kinetochore (inset: microtubules, white; centromere red). A detailed description of the criteria used for scoring lagging chromosomes can be found in Supplementary Fig. 7. Scale bar, 10 μm.

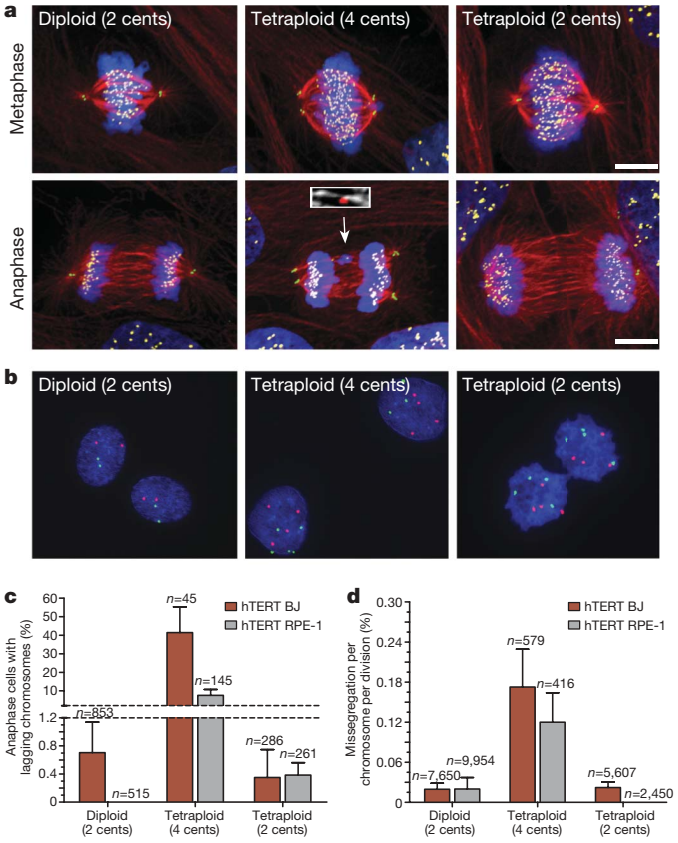


Figure 3 | Extra centrosomes promote chromosome missegregation. **a**, Human hTERT BJ fibroblasts (diploid, two centrosomes; tetraploid, four centrosomes; tetraploid, two centrosomes) during metaphase and anaphase stained for centrioles (green), microtubules (red), chromosomes (blue) and centromeres (yellow). Arrow indicates a lagging chromosome caused by merotelic attachment (inset: microtubules, white; centromere, red). **b**, FISH using centromeric probes specific for chromosomes 6 (green) and 8 (red) in hTERT-BJ fibroblasts. **c**, Percentage of hTERT BJ (red) and hTERT RPE-1 (grey) cells that exhibit one or more lagging chromosomes during bipolar anaphase (*n*, number of anaphases counted). **d**, Missegregation frequency per chromosome per division in hTERT BJ (red) and hTERT RPE-1 (grey) cells (*n*, number of cell divisions counted). Error bars represent mean and s.e.m. from at least 4 independent experiments. Scale bars, 10 μm.

not shown). We therefore used sequential fluorescence-activated cell sorting (FACS) to isolate pure populations of tetraploid cells with a normal complement of centrosomes. Fluorescence *in situ* hybridization (FISH) and karyotyping demonstrated that these cells contained a tetraploid complement of chromosomes (Fig. 3b and Supplementary Fig. 5). Consequently, this procedure generated pure populations of tetraploids with only two centrosomes during mitosis and allowed us to compare the rate of lagging chromosomes in tetraploid cells possessing two or more than two centrosomes. Strikingly, the loss of extra centrosomes was accompanied by a decrease in the fraction of cells with lagging chromosomes to a level observed in diploid cells (Fig. 3a, c). These findings strongly indicate that the increased rate of lagging chromosomes in newly generated tetraploid cells is due to extra centrosomes rather than a duplicated genome.

To determine whether the observed increases in lagging chromosomes in cells containing extra centrosomes leads to chromosome missegregation, we used anaphase/telophase FISH to measure the rate of chromosome missegregation in BJ fibroblasts and RPE-1 cells (Fig. 3b, d). To ensure that cells with extra centrosomes passed through a bipolar mitosis, FISH signals were only scored in anaphase or telophase cells in which the daughters each possessed a single nucleus. The missegregation rates per chromosome in diploid, newly generated tetraploid and late-passage tetraploid cells closely mirrored the results obtained by scoring lagging chromosomes as tetraploids with more than two centrosomes showed missegregation rates of approximately sixfold to eightfold higher than diploids or tetraploids with two centrosomes (Fig. 3d). These rates correspond to one chromosome missegregation for approximately every 6 divisions in tetraploid BJ fibroblasts with more than two centrosomes, compared to one chromosome missegregation for every ~50 divisions in tetraploids with two centrosomes. Thus, extra centrosomes promote chromosome missegregation even after cells cluster centrosomes to assemble bipolar spindles.

Although the above data indicate that most missegregation events can be explained by merotelic attachments and lagging chromosomes, some missegregation events in cells with extra centrosomes may, at low frequency, arise by other mechanisms. For example, we occasionally observed single chromosomes bi-orienting between two inefficiently clustered centrosomes even after all other chromosomes had aligned at the metaphase plate (see Fig. 3a, top row, middle panel). Presumably, these bi-oriented polar chromosomes could be

under tension, satisfy the spindle assembly checkpoint, and thus segregate both sisters to a single daughter upon entry into anaphase [4]. Indeed, in tetraploid RPE-1 cells, we did identify a single such example by live-cell imaging (Supplementary Fig. 6 and Supplementary Movie 3). However, this mechanism does not occur frequently enough to contribute significantly to chromosome missegregation in the extra-centrosomal cells we examined: we did not observe a single such bi-oriented chromosome during anaphase in any of our fixed-cell samples, and only a very minor fraction of the ~8,000 CIN cell divisions we imaged by H2B–GFP showed chromosomes at the poles during anaphase onset, consistent with previous imaging analyses[13,26].

Finally, we designed an experiment in which, in otherwise genetically identical cells, we could directly test the hypothesis that transient multipolar spindle intermediates generate anaphase lagging chromosomes. We recognized that this could be accomplished by monitoring mitosis over two generations after induction of PLK4, a kinase that regulates centriole replication and the overexpression of which has previously been shown to cause centriole amplification[27]. To do this, we used an U2OS osteosarcoma cell line in which PLK4 expression is regulated by a doxycycline-inducible promoter[27]. After a 15 h induction of PLK4, cells contained two 'rosettes' of overduplicated centrioles, each comprised of a mother centriole surrounded by numerous daughter centrioles[27] (Fig. 4a). Importantly, because the extra centrioles assembled before mitosis, they remained engaged and functioned as single units to enable bipolar spindle assembly without a multipolar spindle intermediate (Fig. 4a, b). In these cells, despite centriole amplification, the frequency of lagging chromosomes was similar to that observed in control cells with two centrosomes (Fig. 4b). In contrast, in the second cell cycle after centriole overduplication, supernumerary centrioles disengaged before duplication[28] and multipolar intermediates were common in the mitosis that followed (Fig. 4a, b). In these cells, a marked increase in the frequency of lagging chromosomes was observed after centrosome clustering and the onset of bipolar anaphase (Fig. 4b). Thus, extra centrosomes force cells to pass through a multipolar spindle intermediate and thereby enhance the formation of merotelic attachments and lagging chromosomes.

We propose that extra centrosomes generate CIN primarily by promoting merotelic kinetochore–microtubule attachments. This is due to the unique spindle geometry that occurs when cells resolve transient multipolar intermediates into bipolar spindles, in a manner broadly similar to what occurs during spindle assembly after nocodazole or monastrol washout[20,23] (Fig. 4c). Moreover, syntelic attachments (both sister kinetochores attached to the same pole) are also expected to be enriched when chromosomes bi-orient between two centrosomes that eventually cluster (Fig. 4c), and these syntelic attachments may promote further merotely (Supplementary Movie 4)[21,23]. Many of these merotelic attachment errors are corrected before anaphase by mechanisms that involve the release or destabilization of inappropriately attached microtubules[20,21,29]; this can be inferred by the marked decrease in the frequency and number of lagging chromosomes observed in cells that enter mitosis from a multipolar configuration rather than a bipolar configuration (Supplementary Fig. 4c) or by directly quantifying the number of

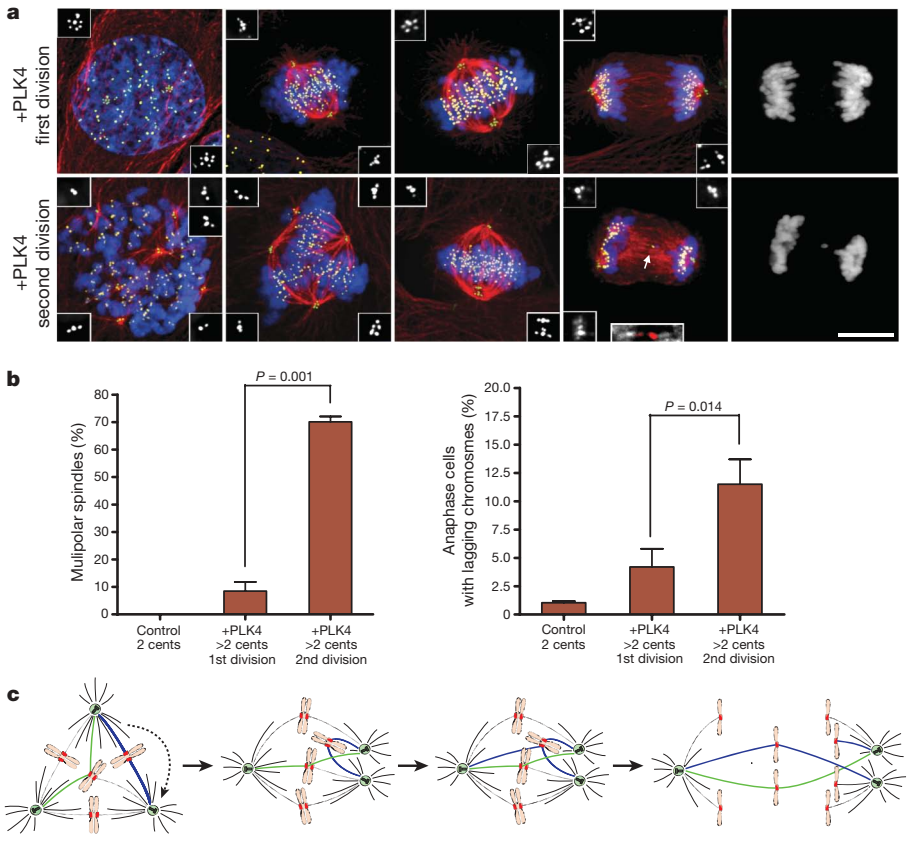


Figure 4 | 'Multipolar spindle intermediates' promote merotelic attachment. **a**, U2OS cells undergoing their first or second mitotic division following PLK4 overexpression stained for centrioles (green; inset, white), microtubules (red), chromosomes (blue) and centromeres (yellow). Arrow indicates a lagging chromosome caused by merotelic attachment (inset: microtubules, white; centromere, red). **b**, The percentage of mitotic cells exhibiting multipolar spindles or lagging chromosomes for each condition. Error bars represent the mean ± s.e.m. from 5 independent experiments; *P*-value derived from paired two-tailed *t*-test. Scale bar, 10 μm. **c**, Extra centrosomes promote merotelic attachment (green microtubules) by altering spindle geometry. In addition, syntelic attachments (blue microtubules) also accumulate upon centrosome clustering and may promote further enhancement of merotely. Unresolved merotelic attachments can give rise to lagging chromosomes at anaphase.

merotelic attachments before and after centrosome clustering (Supplementary Fig. 8). However, the overall increase in the number of initial merotelic attachments during the multipolar spindle intermediate reduces the likelihood that all errors will be corrected before anaphase onset, thereby causing a net increase in the frequency of lagging chromosomes and chromosome missegregation errors[13].

In summary, we have demonstrated that extra centrosomes are not simply innocent bystanders in CIN cells; instead, their presence directly promotes chromosome missegregation that may then facilitate the evolution of more malignant phenotypes. This finding clarifies the longstanding correlation between centrosome amplification and CIN, and provides one simple and unifying explanation for the observed high rates of merotely in CIN cancers. Previously, a variety of genetic mutations have been implicated as causes of CIN, but these defects appear to be relatively infrequent[30]. In contrast, extra centrosomes are prevalent among CIN cells, and we suggest that the mechanism described here is a common contributor to chromosomal instability in human cancer.

METHODS SUMMARY

All cell lines were maintained at 37 °C with 5% CO_2 atmosphere. Immunofluorescence microscopy was performed as previously described[17]. Fixed-cell images were collected by confocal immunofluorescence on a Yokogawa CSU-X1 spinning disk confocal mounted on a Nikon Ti-E inverted microscope (Nikon Instruments). Live-cell imaging was performed using a TE2000-E2 inverted Nikon microscope equipped with the Nikon Perfect Focus system enclosed within a temperature- and CO_2-controlled environment that maintained an atmosphere of 37 °C and 3–5% humidified CO_2. Sequential FACS sorting of tetraploids with 8c DNA content (indicative of tetraploids in the G2/M phase of the cell cycle) was used to generate tetraploid cells with two centrosomes. Detailed descriptions of FISH, karyotyping, imaging, cell lines, culture conditions and antibodies used in this study can be found in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 6 February; accepted 18 May 2009.
Published online 7 June 2009.

1. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instability in colorectal cancers. *Nature* **386,** 623–627 (1997).
2. D'Assoro, A. B., Lingle, W. L. & Salisbury, J. L. Centrosome amplification and the development of cancer. *Oncogene* **21,** 6146–6153 (2002).
3. Nigg, E. A. Centrosome aberrations: cause or consequence of cancer progression? *Nature Rev. Cancer* **2,** 815–825 (2002).
4. Sluder, G. & Nordberg, J. J. The good, the bad and the ugly: the practical consequences of centrosome amplification. *Curr. Opin. Cell Biol.* **16,** 49–54 (2004).
5. Rajagopalan, H. & Lengauer, C. Aneuploidy and cancer. *Nature* **432,** 338–341 (2004).
6. Weaver, B. A., Silk, A. D., Montagna, C., Verdier-Pinard, P. & Cleveland, D. W. Aneuploidy acts both oncogenically and as a tumor suppressor. *Cancer Cell* **11,** 25–36 (2007).
7. Gao, C. *et al.* Chromosome instability, chromosome transcriptome, and clonal evolution of tumor cell populations. *Proc. Natl Acad. Sci. USA* **104,** 8995–9000 (2007).
8. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
9. Ghadimi, B. M. *et al.* Centrosome amplification and instability occurs exclusively in aneuploid, but not in diploid colorectal cancer cell lines, and correlates with numerical chromosomal aberrations. *Genes Chromosom. Cancer* **27,** 183–190 (2000).
10. Lingle, W. L. *et al.* Centrosome amplification drives chromosomal instability in breast tumor development. *Proc. Natl Acad. Sci. USA* **99,** 1978–1983 (2002).
11. Pihan, G. A. *et al.* Centrosome defects can account for cellular and genetic changes that characterize prostate cancer progression. *Cancer Res.* **61,** 2212–2219 (2001).
12. Sato, N. *et al.* Correlation between centrosome abnormalities and chromosomal instability in human pancreatic cancer cells. *Cancer Genet. Cytogenet.* **126,** 13–19 (2001).
13. Thompson, S. L. & Compton, D. A. Examining the link between chromosomal instability and aneuploidy in human cells. *J. Cell Biol.* **180,** 665–672 (2008).
14. Torres, E. M., Williams, B. R. & Amon, A. Aneuploidy: cells losing their balance. *Genetics* **179,** 737–746 (2008).
15. Brinkley, B. R. Managing the centrosome numbers game: from chaos to stability in cancer cell division. *Trends Cell Biol.* **11,** 18–21 (2001).
16. Basto, R. *et al.* Centrosome amplification can initiate tumorigenesis in flies. *Cell* **133,** 1032–1042 (2008).
17. Kwon, M. *et al.* Mechanisms to suppress multipolar divisions in cancer cells with extra centrosomes. *Genes Dev.* **22,** 2189–2203 (2008).
18. Quintyne, N. J., Reing, J. E., Hoffelder, D. R., Gollin, S. M. & Saunders, W. S. Spindle multipolarity is prevented by centrosomal clustering. *Science* **307,** 127–129 (2005).
19. Yang, Z., Loncarek, J., Khodjakov, A. & Rieder, C. L. Extra centrosomes and/or chromosomes prolong mitosis in human cells. *Nature Cell Biol.* **10,** 748–751 (2008).
20. Cimini, D. Merotelic kinetochore orientation, aneuploidy, and cancer. *Biochim. Biophys. Acta* **1786,** 32–40 (2008).
21. Salmon, E. D., Cimini, D., Cameron, L. A. & DeLuca, J. G. Merotelic kinetochores in mammalian tissue cells. *Phil. Trans. R. Soc. Lond. B* **360,** 553–568 (2005).
22. Cimini, D. *et al.* Merotelic kinetochore orientation is a major mechanism of aneuploidy in mitotic mammalian tissue cells. *J. Cell Biol.* **153,** 517–527 (2001).
23. Cimini, D., Moree, B., Canman, J. C. & Salmon, E. D. Merotelic kinetochore orientation occurs frequently during early mitosis in mammalian tissue cells and error correction is achieved by two different mechanisms. *J. Cell Sci.* **116,** 4213–4225 (2003).
24. Heneen, W. K. Kinetochores and microtubules in multipolar mitosis and chromosome orientation. *Exp. Cell Res.* **91,** 57–62 (1975).
25. Sluder, G., Thompson, E. A., Miller, F. J., Hayes, J. & Rieder, C. L. The checkpoint control for anaphase onset does not monitor excess numbers of spindle poles or bipolar spindle symmetry. *J. Cell Sci.* **110,** 421–429 (1997).
26. Gascoigne, K. E. & Taylor, S. S. Cancer cells display profound intra- and interline variation following prolonged exposure to antimitotic drugs. *Cancer Cell* **14,** 111–122 (2008).
27. Kleylein-Sohn, J. *et al.* Plk4-induced centriole biogenesis in human cells. *Dev. Cell* **13,** 190–202 (2007).
28. Tsou, M. F. & Stearns, T. Mechanism limiting centrosome duplication to once per cell cycle. *Nature* **442,** 947–951 (2006).
29. Bakhoum, S. F., Thompson, S. L., Manning, A. L. & Compton, D. A. Genome stability is ensured by temporal control of kinetochore-microtubule dynamics. *Nature Cell Biol.* **11,** 27–35 (2008).
30. Chandhok, N. S. & Pellman, D. A little CIN may cost a lot: revisiting aneuploidy and cancer. *Curr. Opin. Genet. Dev.* **19,** 74–81 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank D. Compton, R. King, D. Livingston, A. Manning, T. Rapoport, J. Shah, T. Stukenberg and members of the Pellman laboratory for comments and discussion of the manuscript; L. Cameron for setting up the confocal microscope; P. Schow for help with FACS; D. Neuberg for help with statistical analysis; C. King for help with the model figure; J. Iwasa for creating the animations; and L. Moreau and S. Thompson for technical advice on chromosome spreads and FISH analysis. Reagents were provided by D. Compton, S. Gollin, R. King, A. Khodjakov, E. Nigg and G.Wahl. N.J.G. is a fellow of the Leukemia and Lymphoma Society and D.P. is an HHMI Investigator. This work was supported by NIH grant GM083299.

Author Contributions N.J.G. and D.P. designed the experiments and wrote the manuscript. N.J.G. performed most of the experiments and data analysis. S.A.G. conceived and conducted experiments and performed data analysis, particularly for Fig. 4. D.P. supervised all aspects of the project. All authors discussed results and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.P. (david_pellman@dfci.harvard.edu).

METHODS

Cell culture. Cell lines were maintained at 37 °C with 5% CO_2 atmosphere in DMEM (U20S, hTERT-BJ, MCF-7, N1E-115, CFPAC-1), McCoy's (Caco2, HT-29, MDA-231), DMEM:F12 (hTERT-RPE1), MEM (UPCI:SCC114) or RPMI (BT549) containing 10% FBS, 100 IU ml^{-1} penicillin and 100 μg ml^{-1} streptomycin. Mouse *p53*$^{-/-}$ tetraploid cells from tumours (tMMECs) were grown in DMEM:F12 supplemented with 2% FBS, 2 mg ml^{-1} insulin, 25 μg ml^{-1} epidermal growth factor, 100 IU ml^{-1} penicillin, 100 μg ml^{-1} streptomycin and 0.25 μg ml^{-1} fungizone. The tetracycline-inducible U2OS cell line expressing MYC-PLK44 (a gift from E. Nigg) was grown in DMEM supplemented with 10% of tetracycline-free FBS, 100 IU ml^{-1} penicillin, 100 μg ml^{-1} streptomycin, 50 μg ml^{-1} hygromycin and 1 mg ml^{-1} G418. Myc-Plk4 expression was induced by addition of 1 μg ml^{-1} doxycyclin for 15 h.

To generate tetraploid cells with extra centrosomes, hTERT BJ and RPE-1 cells were transfected with 50 nmol p53 siRNA (SmartPool, Dharmacon) using Lipofectamine RNAi Max (Invitrogen). Approximately 6 h later, fresh medium containing 0.2 μm cytochlasin D was added for an additional 16 h. Cells were then washed six times for 5 min each to remove cytochalasin D and allowed to recover for ~6–24 h.

Indirect immunofluorescence microscopy. All cells stained for centrioles were washed in PBS and then fixed in ice-cold methanol for 10 min. After fixation, cells were rehydrated in PBS for 5 min, extracted in PBS-0.5% Triton X-100 for 5 min, blocked in TBS-BSA (10 mM Tris, pH 7.5, 150 mM NaCl, 5% BSA, 0.1% Tween) for 30 min, and then incubated with primary antibodies in TBS-BSA for 30–60 min. Antibodies included mouse anti-tubulin (DM1α, 1:500; Sigma-Aldrich), human anti-centromere ACA (1:500; Antibodies, Inc.) and rabbit anti-centrin2 (1:200; Santa Cruz).

To visualize merotelic kinetochores, cells were permeabilized in Ca^{2+} buffer (100 mM PIPES, pH 6.8, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 0.5% Triton X-100) for 5–10 min at room temperature (20–23 °C), followed by fixation in 1% glutaraldehyde in PBS for 10 min, and two 12-min washes in freshly prepared 0.5 mg ml^{-1} $NaBH_4$. After a 60-min block in TBS-BSA, cells were incubated with primary antibodies in TBS-BSA for 2–4 h. Antibodies included rat anti-tubulin (YL 1/2, 1:1,000; Novus), rabbit anti-pericentrin (1:1,000; AbCam) and mouse anti-Hec1 (1:200, Novus). All primary antibodies were detected using species-specific fluorescent secondary antibodies (Molecular Probes) and DNA was detected with 0.2 μg ml^{-1} 4,6-diamidino-2-phenylindole (DAPI; Sigma-Aldrich). Coverslips were mounted with ProLong Antifade mounting medium (Molecular Probes).

Confocal immunofluorescence images were collected at 405, 488, 561 and 640 nm with a Yokogawa CSU-X1 spinning disk confocal mounted on a Nikon Ti-E inverted microscope (Nikon Instruments). A series of 0.2-μm optical sections were acquired using a ×100 1.4 NA Plan Apo objective lens with an Orca ER CCD camera (Hamamatsu Photonics). Acquisition parameters, shutters, filter positions and focus were controlled by Andor iQ software. Images presented in figures are maximum intensity projections of entire *z*-stacks, except for high-magnification images of merotelic attachments, which are comprised of only enough *z*-layers to visualize entire kinetochores. Cells were classified as having more than two centrosomes if more than two distinct pairs of centrioles were observed during mitosis. Lagging chromosomes were scored as centromere-positive chromosomes that were completely separated from the two main masses of chromosomes during mid-anaphase (see Supplementary Fig. 7). Kinetochores were scored as being merotelically attached if they were clearly attached to two visible kinetochore fibres emanating from opposite poles.

Long-term live-cell imaging. H2B–GFP was cloned into the pLenti6/V5 lentiviral vector (Invitrogen). Cells infected with lentivirus encoding H2B–GFP were FACS-sorted by GFP fluorescence. GFP-positive cells were grown on glass-bottomed 12-well tissue culture dishes (Mattek) and imaged on a TE2000-E2 inverted Nikon microscope equipped with a cooled CCD camera (Orca ER, Hamamatsu) and the Nikon Perfect Focus system. The microscope was enclosed within a temperature- and CO_2-controlled environment that maintained an atmosphere of 37 °C and 3–5% humidified CO_2. GFP and/or differential interference contrast images were captured at multiple points every 10 min for 3–5 days with a ×20 0.5 NA Plan Fluor objective. All captured images were analysed using NIS-Elements software.

Short-term live-cell imaging. GFP–centrin RPE-1 cells (a gift from A. Khodjakov) were infected with retrovirus encoding monomeric red fluorescent protein (mRFP)–H2B and sorted by RFP fluorescence. For imaging, cells were grown on glass-bottomed 12-well tissue culture dishes (Mattek) overlayed with mineral oil and maintained at 37 °C. Confocal immunofluorescence images were collected at 488 and 561 nm with a Yokogawa CSU-X1 spinning disk confocal mounted on a Nikon Ti-E inverted microscope (Nikon Instruments). A series of 1 μm optical sections were acquired every 2 min using a ×60 1.4 NA Plan Apo objective lens with an Orca ER CCD camera (Hamamatsu Photonics). Images presented in figures are maximum intensity projections of entire *z*-stacks.

Generation of tetraploid cells with two centrosomes. Human hTERT RPE-1 and hTERT BJ cells were treated with 0.2 μm cytochalasin D for ~16 h, washed every 5 min for 30 min, and then FACS-sorted by DNA content using Hoechst at 1:2,500 (Molecular Probes). Cells with a DNA content of 8c were isolated and cultured for ~1 week before a second FACS sorting to re-isolate 8c cells. A significant portion (~50%) of tetraploid cells displayed two centrosomes after only two sorts, and these cells were used to measure lagging chromosomes. By sort four, nearly 100% of tetraploid cells had two centrosomes, and these were used for FISH analysis. We verified p53 function was intact in these late-passage tetraploid BJ and RPE-1 cells by western blotting for phospho-p53 (Ser 15) after 5 h treatment with 200 ng ml^{-1} doxorubicin.

Chromosome spreads. Diploid and late-passage tetraploid cells were treated with colcemid (50 ng ml^{-1}) for 4 h, trypsinized and resuspended in 0.56% KCl for 30 min at 37 °C. Cells were then fixed with 3:1 ice-cold methanol:acetic acid, pelleted, and then washed twice more with methanol:acetic acid before being dropped on a pre-cleaned glass side from a height of ~12 inches. Cells were allowed to dry on the slide and were then stained for 3 min with Giemsa stain in 1× Gurr's buffer (Gibco). After a wash in Gurr's buffer, coverslips were added to slides and sealed with Permount (Fisher). Images of spreads were taken with a ×100 objective on a Zeiss upright microscope and chromosomes were counted using PhotoShop.

FISH. Diploid and late-passage hTERT BJ and RPE-1 cells were transfected with 50 nmol p53 siRNA (SmartPool, Dharmacon) using Lipofectamine RNAi Max (Invitrogen). Approximately 6 h later, fresh medium containing 0.2 μm cytochlasin D was added for an additional 16 h. Cells were then washed six times for 5 min each to remove cytochalasin D and allowed to recover for ~6–10 h. Mitotic shake-off was then used to isolate dividing cells at a low density on fresh coverslips. Once cells were reattached, 0.56% KCl was added to the coverslips for 30 min followed by fixation in 3:1 methanol acetic acid. After two additional methanol:acetic acid washes, coverslips were dried completely and aged for two days. Coverslips were then washed in 2× SSC plus 0.5% NP-40 for 30 min at 37 °C and then dehydrated by sequential 2-min washes in 70%, 85% and 100% ethanol. Fluorescently labelled centromeric FISH probes specific for chromosomes 6, 8, 7 or 11 (Cytocell) were diluted 1:10 into hybridization buffer (Cytocell) and added to coverlips. Coverlips were affixed to a pre-cleaned slide with rubber cement and then placed in a hybridization oven at 75 °C for 5 min, followed by an overnight incubation at 37 °C in a humidified chamber. The next day, cells were washed in 1× PBD (0.1 M NaH_2PO_4, 0.1 M Na_2HPO_4, 0.1% NP-40) for 2 min to remove the rubber cement followed by a wash in 1× wash buffer (0.5× SSC plus 0.1% SDS) for 5 min at 65 °C. Coverslips were then incubated with 1× PBD plus 0.2 μg ml^{-1} DAPI for 20 min before being sealed on a slide with mounting medium. FISH signals were only counted in daughter cells that possessed a single nucleus, eliminating the possibility that the cells had undergone a multipolar mitosis followed by cytokinesis failure. FISH signals in micronuclei were ignored because it could not be definitively determined which of the two daughter cells the micronuclei were within. Thus, our rates of chromosome missegregation in tetraploid cells with extra centrosomes are likely to be underestimates. The reported rates of chromosome missegregation are a sum of each of the four individual probes we used (6, 8, 7, 11).

LETTERS

Uniparental expression of PolIV-dependent siRNAs in developing endosperm of *Arabidopsis*

Rebecca A. Mosher[1], Charles W. Melnyk[1], Krystyna A. Kelly[1], Ruth M. Dunn[1], David J. Studholme[2] & David C. Baulcombe[1]

Most eukaryotes produce small RNA (sRNA) mediators of gene silencing that bind to Argonaute proteins and guide them, by base pairing, to an RNA target. MicroRNAs (miRNAs) that normally target messenger RNAs for degradation or translational arrest are the best-understood class of sRNAs. However, in *Arabidopsis thaliana* flowers, miRNAs account for only 5% of the sRNA mass and less than 0.1% of the sequence complexity. The remaining sRNAs form a complex population of more than 100,000 different small interfering RNAs (siRNAs) transcribed from thousands of loci[1–5]. The biogenesis of most of the siRNAs in *Arabidopsis* are dependent on RNA polymerase IV (PolIV), a homologue of DNA-dependent RNA polymerase II[2,3,6]. A subset of these PolIV-dependent (p4)-siRNAs are involved in stress responses, and others are associated with epigenetic modifications to DNA or chromatin; however, the biological role is not known for most of them. Here we show that the predominant phase of p4-siRNA accumulation is initiated in the maternal gametophyte and continues during seed development. Expression of p4-siRNAs in developing endosperm is specifically from maternal chromosomes. Our results provide the first evidence for a link between genomic imprinting and RNA silencing in plants.

Arabidopsis plants in which the largest subunit of PolIV (NRPD1A) is either absent or non-functional have no obvious morphological or fertility defects. This absence of strong phenotypes is surprising because the p4-siRNAs are produced from thousands of loci corresponding to more than 1% of the genome[3,6]. In *nrpd1a* mutants there is decreased DNA methylation and reactivation of transcription at some transposable elements[7–10], which is consistent with a role of p4-siRNAs in genome defence. However, a straightforward role in genome defence is unlikely because the *NRPD1A* mutant phenotype is unlike that of the chromatin-remodelling and genome defence protein DDM1 (refs 7, 9). The *ddm1* mutants show a progressive increase in morphological and growth defects over several generations as a result of transposon mobilization[11,12], whereas the equivalent *nrpd1a* plants grow normally.

To explore the role of p4-siRNAs we monitored their levels in different tissues of *Arabidopsis*. Consistent with previous observations that sRNAs are most abundant in floral tissues[2,13,14] was our finding that a subset of p4-siRNAs are present only in flowers and young siliques, whereas a second class of p4-siRNAs could be detected throughout development (Fig. 1a and Supplementary Fig. 1). Floral-specific p4-siRNAs (type I) were additionally characterized by a burst of expression in young siliques. The levels of these p4-siRNAs peaked about 4–6 days after anthesis (mid-embryogenesis) before returning to undetectable levels in mature siliques (Fig. 1b, c). This burst of p4-siRNA production was from the developing seeds rather than the valves or septum of the silique (Fig. 1d) and was primarily from the endosperm or seed coat of the developing seed (Fig. 1e). On the basis of the intensity of northern blotting signal, we estimate that the level of these type I p4-siRNAs in the endosperm or seed coat is higher than at any stage in the life cycle of the plant. Although the type II p4-siRNAs are present in vegetative tissues, including the developing embryo, they were also found predominantly in developing endosperm or seed coat.

To investigate the relationship between the expression of p4-siRNAs before and after fertilization, we performed crosses between wild-type and *nrpd1a-4* plants. In siliques from an *nrpd1a-4* (female) × *NRPD1A* (male) cross, both type I and type II p4-siRNAs were undetectable (Fig. 2a), indicating that paternal *NRPD1A* does not contribute to p4-siRNA expression in the seed. Similar crosses show that paternal *RDR2* and *DCL3* are also insufficient for p4-siRNA accumulation during seed development (see Supplementary Information).

The possibility that p4-siRNAs after fertilization are produced in the maternally derived seed coat can be ruled out on the basis of our analysis of unfertilized flowers of an *NRPD1A*/*nrpd1a-4* heterozygote. If these RNAs were produced in diploid maternal tissue they would be as abundant in the flowers of the heterozygote as in the wild type. However, the p4-siRNAs were significantly less abundant than in wild-type plants (Fig. 2b and Supplementary Fig. 2), which is consistent with the alternative possibility that they are produced in the 50% of the female gametophytes carrying the maternal *NRPD1A* allele (see Supplementary Information). From this analysis of the progeny of *NRPD1A*/*nrpd1a-4* heterozygotes it is likely that the burst of p4-siRNA expression is initiated in the female gametophyte and persists in either the embryo or the endosperm. Further analysis of dissected seeds (Fig. 1e) indicated that the massive expression of p4-siRNAs after fertilization is predominantly from the endosperm.

The expression pattern of p4-siRNAs may indicate maternal-specific expression of either p4-siRNA biosynthetic genes or of the genomic loci that specify these sRNAs. To investigate the first possibility we analysed mRNAs from reciprocal crosses of *Arabidopsis* genotypes Columbia-0 (Col) and Landsberg *erecta* (Ler) in developing seeds. With an RT–PCR cleaved amplified polymorphic sequence (CAPS) analysis we could detect both maternal and paternal transcripts of the p4-siRNA biosynthetic genes *NRPD1A*, *DCL3* and *RDR2* in these samples, indicating that they are biparentally expressed (Fig. 2c and Supplementary Fig. 3). To test the second hypothesis we used genotype-specific oligonucleotide probes to assay p4-siRNAs from locus 08002 in a Col × Ler cross. Northern analysis with these probes confirmed the expression of the maternal but not the paternal locus 08002 in the developing seed (Fig. 3a). From this analysis we conclude that the uniparental expression is due to an effect at maternal loci that specify sRNA and not at loci encoding proteins in the siRNA pathways. The contrasting expression patterns of p4-siRNAs and of *NRPD1A*, *DCL3*

[1]Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK. [2]The Sainsbury Laboratory, Colney Lane, Norwich NR4 7UH, UK.

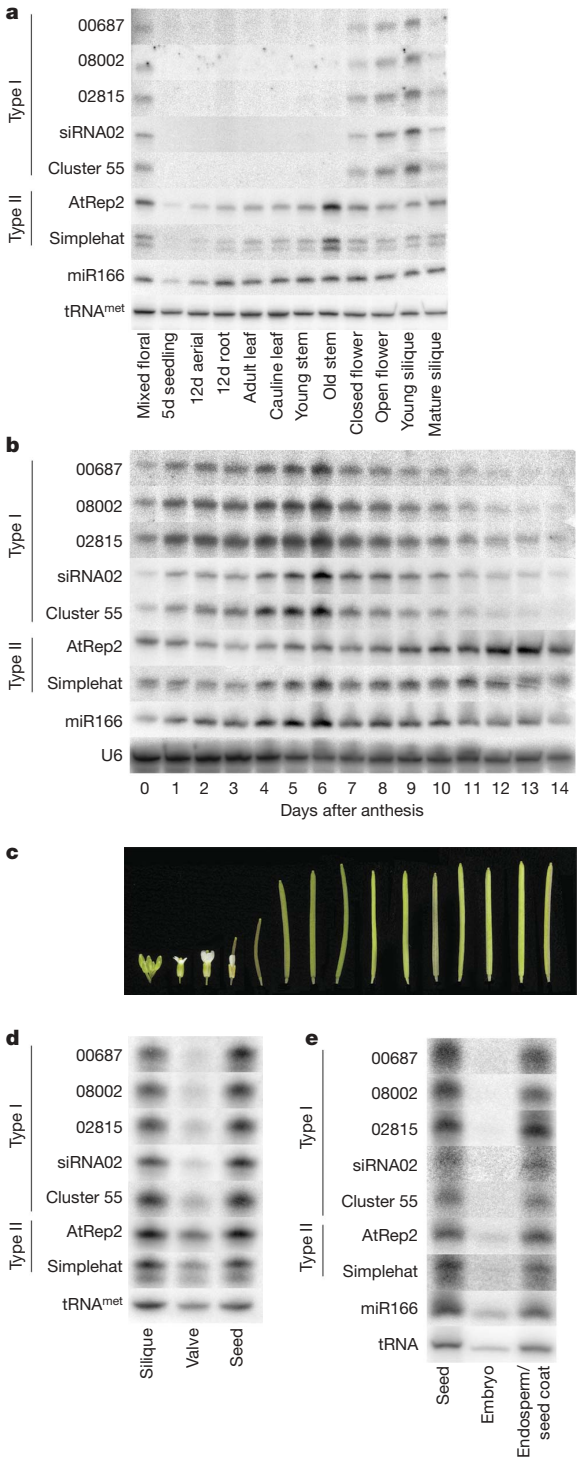


Figure 1 | p4-siRNAs accumulate to high concentrations in seeds during early embryogenesis. **a**, Various plant tissues were assayed by northern blotting for the presence of p4-siRNAs. Type I p4-siRNAs accumulate specifically in flowers and young fruits, whereas type II p4-siRNAs are also expressed throughout development, more like a microRNA (miR166); $tRNA^{met}$, loading control. **b**, p4-siRNAs are present at 0 days after anthesis (before fertilization) and are abundant four to six days after anthesis (early embryogenesis). U6, loading control. **c**, Representative flowers and siliques at the noted day after anthesis. **d**, Young siliques were manually dissected to isolate developing seeds from fruit tissues. RNA from ten whole siliques or ten dissected siliques was loaded in each lane for northern blotting. p4-siRNA expression is mainly in developing seeds (seed) rather than the valves and septum of the fruit (valve). **e**, Immature seeds were dissected into embryo and endosperm/seed-coat fractions before RNA extraction. RNA from all of the seeds within two siliques was run in each lane. $tRNA^{met}$, loading control. Type I p4-siRNAs are abundant only in the endosperm/seed-coat fraction, whereas type II p4-siRNAs accumulate to lower levels in the embryo, similar to those of miR166.

and *RDR2* also provide clear evidence that the uniparental expression of the p4-siRNA is distinct from the delayed activation of the paternal genome in seeds (see Supplementary Information)[15].

Next, to find out whether maternal expression of p4-siRNAs is a genome-wide phenomenon, we deep-sequenced sRNAs from siliques of reciprocal Col × Ler crosses and analysed the 24 nucleotide reads characteristic of p4-siRNAs (Fig. 3b and Supplementary Table 1). Those reads perfectly matching only the Ler genome correspond to Ler-specific sequence motifs in the sRNA data sets. However, the Col-specific motifs could not be identified precisely because the available Ler genome sequence is incomplete. From this analysis we cannot exclude the possibility that a minority of p4-siRNA loci are biparentally expressed. However, the presence of the Ler paternal genome does not greatly alter the representation of Ler-specific sequences (see Supplementary Information), which is consistent with maternal-specific expression of all or most of the p4-siRNA loci. In contrast, the PolIV-independent *trans*-acting siRNAs from both Col and Ler alleles of the *TAS1A* locus are represented in the developing seed data sets (Supplementary Fig. 4) and are therefore biparentally expressed.

Maternal-specific expression has previously been described for *FIS2* and *FWA* protein-coding genes. To find out whether the same mechanism is involved with these genes and with p4-siRNA loci we used *Arabidopsis* lines carrying *met1-1*. *MET1* is required for the repression of *FIS2* and *FWA* in the male gametophyte, and fertilization with *met1* pollen results in biparental expression of these genes[16,17]. In our analyses, as expected, paternal *FWA* transcripts were detected in the progeny of crosses between Ler and *met1-1* (Col genotype). However, the paternal p4-siRNA locus 08002 was not activated in this cross (Fig. 3c, d). We can therefore rule out the possibility that *FIS2*/*FWA* and the p4-siRNA loci are suppressed in the paternal lineage by the same mechanism.

To investigate the connection between maternal expression of p4-siRNAs in the developing seed and p4-siRNA accumulation in the resulting adult plant we monitored type II p4-siRNAs and p4-siRNA-dependent DNA methylation in leaves of *nrpd1a-4*/*NRPD1A* heterozygotes. These data confirm the presence of p4-siRNAs in leaves. In addition, at least for the *AtSN1* locus that is a target of siRNA-directed methylation, the heterozygous mutations have no effect in leaves: siRNA-directed DNA methylation was at wild-type levels regardless of the presence of p4-siRNAs during seed development (Fig. 4a, b). Similarly, in the flowers of Col × Ler F_1 hybrid plants, there were locus 08002 p4-siRNAs from both Col and Ler alleles (Fig. 4c). It is therefore likely that any epigenetic marks responsible for uniparental expression of p4-siRNA in developing seeds are lost as the embryo develops into the mature plant.

This discovery of maternally expressed p4-siRNAs in developing seeds greatly expands the catalogue of imprinted loci in the *Arabidopsis* genome. Previously, four imprinted loci were known in *Arabidopsis*, all of which encoded proteins[17–20]. Now, with the demonstration of maternal expression of p4 siRNA loci, we produce evidence for an imprinting phenomenon that is more widely distributed through the *Arabidopsis* genome than the uniparental expression pattern in any other organism.

The uniparental expression pattern of p4-siRNAs requires that there be either an activating factor produced in the maternal lineage, a repressive factor in the paternal lineage, or both. The data produced here are consistent with an activating factor linked to maternal expression of *NRPD1A*, *RDR2* and *DCL3*: loss of function of these genes in the maternal lineage prevents the expression of p4-siRNAs. To explain the effect of the activating factor on the maternal p4-siRNA loci we propose that it acts in the female gametophyte rather than after fertilization. The activators would impose an epigenetic

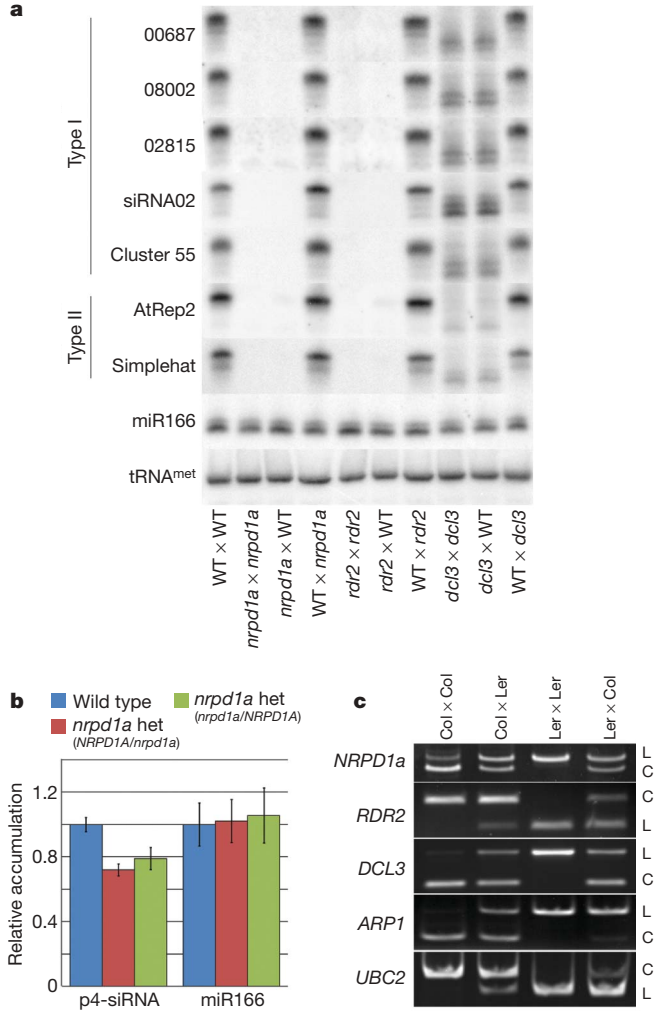


Figure 2 | p4-siRNA expression in the zygote requires maternal NRPD1A. **a**, Emasculated flowers were pollinated manually and RNA was extracted from the developing cross at five days after fertilization. The 24-nt p4-siRNAs accumulate in developing seeds only when p4-siRNA biosynthetic genes are inherited maternally. miR166 does not require PolIV biosynthetic genes. WT, wild type; $tRNA^{met}$, loading control. **b**, p4-siRNA accumulation in closed flowers from wild-type and *nrpd1a* heterozygotes (*NRPD1A*/*nrpd1a* or *nrpd1a*/*NRPD1A*) was quantified with northern blots and image-processing software. Three biological replicates were assayed at four different p4-siRNA loci. Error bars indicate s.e.m. In an analysis of variance, accumulation of p4-siRNAs in *nrpd1a-4* heterozygotes is significantly lower than in the wild type ($F_{1,48} = 37.8$; $P < 10^{-7}$), indicating a gametophytic requirement for NRPD1A. **c**, RNA was extracted from dissected seeds of Columbia-0 (Col) × Landsberg *erecta* (Ler) crosses at five days after fertilization. RT–PCR CAPS analysis indicates that paternal alleles of PolIV biosynthetic loci (*NRPD1A*, *RDR2* and *DCL3*) are expressed at time points when maternal alleles are required for p4-siRNA biosynthesis. *ARP1* and *UBC2* are control genes. L, Ler allele; C, Col allele.

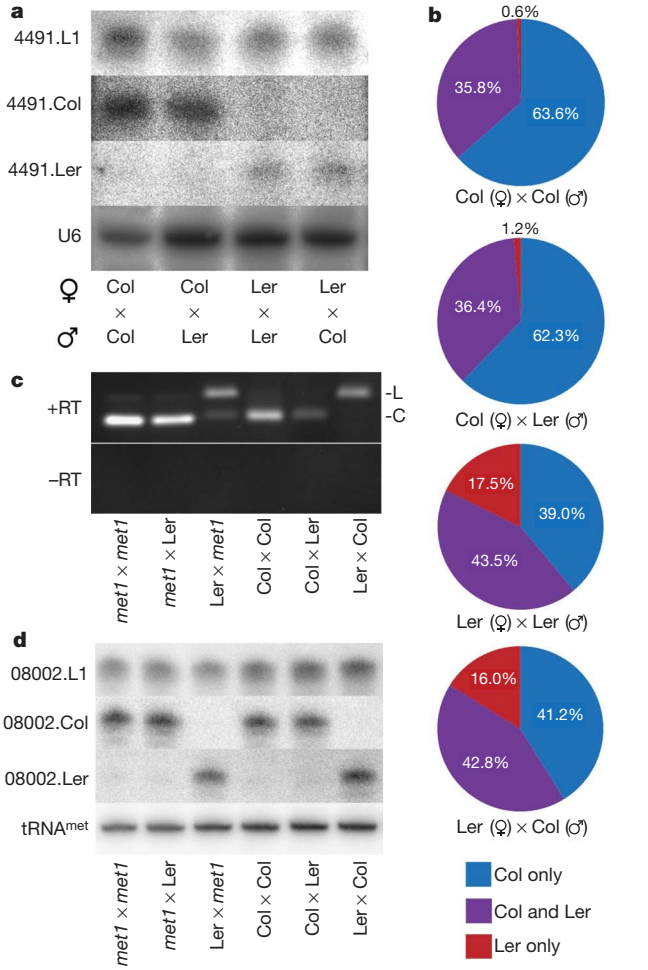


Figure 3 | p4-siRNA expression in developing seeds is uniparental genome-wide. **a**, RNA was purified from crosses between genotypes Columbia-0 (Col) and Landsberg *erecta* (Ler). RNA samples were taken at five days after fertilization and used for northern blot analysis. Allele-specific probes to the p4-siRNA locus 08002 (08002.Col and 08002.Ler) detect only expression from the maternal allele in each cross. Probe 08002.L1 binds both alleles. U6, loading control. **b**, Small RNA libraries were prepared from tissues as in **a** and deep-sequenced. The frequency of 24-nt sRNAs matching to the Col or Ler genome is unaffected by paternal genotype, indicating that the paternal genome is not a source of p4-siRNAs. **c**, RNA was purified from crosses between *met1-1* (genotype Col) and Landsberg *erecta* (Ler) at five days after fertilization and allele-specific FWA transcripts were detected by RT–PCR CAPS. When *met1-1* was the pollen donor, Col-specific transcripts (C) are detected, indicating biparental expression. **d**, Northern blot analysis was performed on sRNA from the same crosses as in **c**. Allele-specific probes to the p4-siRNA locus 08002 (08002.Col and 08002.Ler) detect only expression from the maternal allele in each cross. Probe 08002.L1 binds a conserved region of this locus and binds both alleles. $tRNA^{met}$, loading control.

mark on the maternal p4-siRNA loci that persists into the endosperm of the fertilized seed. However, a repressive paternal factor may also be involved. An absence of p4-siRNAs derived from transposable elements in the sperm cells of *Arabidopsis* has been reported[21]. The loss of p4-siRNA production in sperm cells is coincident with a loss of the chromatin-remodelling protein DDM1, indicating that chromatin structure may repress the expression of paternal p4-siRNA after fertilization.

It is likely that the maternally expressed p4-siRNAs can affect gene expression by acting *in trans* at either the transcriptional level or the post-transcriptional level. However, *nrpd1a* mutant genotypes do not have a strong growth phenotype. To resolve this paradox we propose that p4-siRNA loci would have been selected against if they silenced essential genes in their genome of origin. However, because the population of p4-siRNAs is so large, there is a good chance that some of them would target genes in a heterologous genome and that they would influence the phenotype of hybrids between genotypes and related species. In effect the p4-siRNA system would allow the differentiation of self genes from non-self genes. Such a role based on a maternal expression pattern could contribute to the observed parental genome conflict in which the maternal genome has a suppressive effect on hybrids. In addition, because the interaction between maternal p4-siRNA and a paternal target is a unique property of hybrids, it could contribute to phenotypes that are not observed in the parents. These phenotypes could include growth

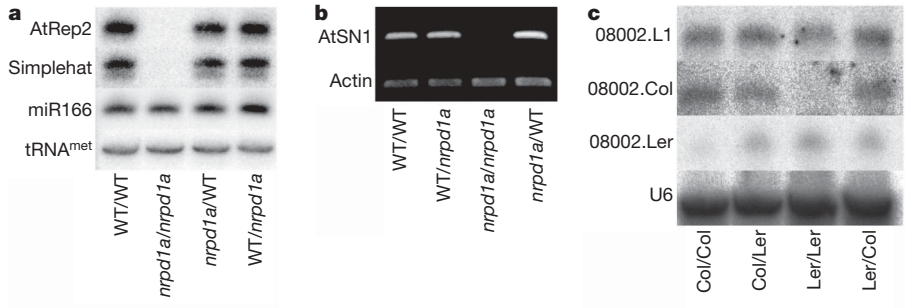


Figure 4 | Uniparental expression of p4-siRNAs is specific to the seed. **a**, **b**, RNA and DNA was extracted from adult leaves of *nrpd1a-4* heterozygotes and the accumulation of p4-siRNAs (**a**) and RNA-directed DNA methylation (**b**) were assayed. Both heterozygotes are equivalent to the wild type (WT), indicating that p4-siRNA accumulation and action occurs in adults regardless of parental history. **c**, RNA was isolated from mixed floral tissue of Col/Ler F_1 hybrids and p4-siRNAs were detected from both alleles of locus 08002. U6, loading control.

enhancement through hybrid vigour and transgressive segregation or suppressive effects that are manifested as a post-zygotic failure. All of these effects could be readily analysable in plants that outcross more frequently than *Arabidopsis*.

METHODS SUMMARY

For all crosses, immature flowers were emasculated 24 h before manual pollination and were collected five days after fertilization or allowed to proceed to maturity. Mutant alleles used were as follows: *nrpd1a-4*, *rdr2-2*, *dcl3-1* and *met1-1*. Standard laboratory techniques were followed for all RNA isolation, reverse transcription, PCR, and enzymatic digestions. For northern blotting, small (less than 200-nucleotide (nt)) RNA was purified from total RNA with a mirVana miRNA isolation kit (Ambion) before separation on a 15% TBE/urea gel and transfer to membrane. Hybridization of labelled oligonucleotides occurred in Ambion UltraHyb Oligo buffer at 35 °C. For quantification of northern blot signals, three biological replicates were measured blind with image-processing software. p4-siRNA or miR166 signals were normalized to $tRNA^{met}$ signals as a loading control. Values for locus 00687, locus 08002, locus 02815 and cluster 55 were combined to generate a 'p4-siRNA' value. Libraries of sRNAs were prepared by 5′ and 3′ linker ligation followed by RT–PCR, and were sequenced on an Illumina Genome Analyser.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 January; accepted 21 April 2009.
Published online 3 June 2009.

1. Aravin, A. A. *et al.* The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* **5,** 337–350 (2003).
2. Kasschau, K. D. *et al.* Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol.* **5,** 0479–0493 (2007).
3. Mosher, R. A., Schwach, F., Studholme, D. J. & Baulcombe, D. C. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc. Natl Acad. Sci. USA* **105,** 3145–3150 (2008).
4. Reinhart, B. J., Weinstein, E. G., Rhoades, M., Bartel, B. & Bartel, D. P. MicroRNAs in plants. *Genes Dev.* **16,** 1616–1626 (2002).
5. Ruby, J. G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional MicroRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127,** 1193–1207 (2006).
6. Zhang, X., Henderson, I. R., Lu, C., Green, P. J. & Jacobsen, S. E. Role of RNA polymerase IV in plant small RNA metabolism. *Proc. Natl Acad. Sci. USA* **104,** 4536–4541 (2007).
7. Herr, A. J., Jensen, M. B., Dalmay, T. & Baulcombe, D. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308,** 118–120 (2005).
8. Kanno, T. *et al.* Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nature Genet.* **37,** 761–765 (2005).
9. Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120,** 613–622 (2005).
10. Pontier, D. *et al.* Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in *Arabidopsis*. *Genes Dev.* **19,** 2030–2040 (2005).
11. Kakutani, T., Jeddeloh, J. A., Flowers, S. K., Munakata, K. & Richards, E. J. Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. *Proc. Natl Acad. Sci. USA* **93,** 12406–12411 (1996).
12. Singer, T., Yordan, C. & Martienssen, R. A. Robertson's *Mutator* transposons in *A. thaliana* are regulated by the chromatin-remodeling gene *Decrease in DNA Methylation* (*DDM1*). *Genes Dev.* **15,** 591–602 (2001).
13. Lu, C. *et al.* Elucidation of the small RNA component of the transcriptome. *Science* **309,** 1567–1569 (2005).
14. Swiezewski, S. *et al.* Small RNA-mediated chromatin silencing directed to the 3′ region of the *Arabidopsis* gene encoding the developmental regulator, FLC. *Proc. Natl Acad. Sci. USA* **104,** 3633–3638 (2007).
15. Vielle-Calzada, J. P., Baskar, R. & Grossniklaus, U. Delayed activation of the paternal genome during seed development. *Nature* **404,** 91–94 (2000).
16. Jullien, P. E., Katz, A., Oliva, M., Ohad, N. & Berger, F. Polycomb group complexes self-regulate imprinting of the Polycomb group gene *MEDEA* in *Arabidopsis*. *Curr. Biol.* **16,** 486–492 (2006).
17. Kinoshita, T. *et al.* One-way control of *FWA* imprinting in *Arabidopsis* endosperm by DNA methylation. *Science* **303,** 521–523 (2004).
18. Jullien, P. E., Kinoshita, T., Ohad, N. & Berger, F. Maintenance of DNA methylation during the *Arabidopsis* life cycle is essential for parental imprinting. *Plant Cell* **18,** 1360–1372 (2006).
19. Kinoshita, T., Yadegari, R., Harada, J. J., Goldberg, R. B. & Fischer, R. L. Imprinting of the *MEDEA* Polycomb gene in the Arabidopsis endosperm. *Plant Cell* **11,** 1945–1952 (1999).
20. Kohler, C. *et al.* The *Polycomb*-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene *PHERES1*. *Genes Dev.* **17,** 1540–1553 (2003).
21. Slotkin, R. K. *et al.* Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136,** 461–472 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank I. Furner for *met1-1* seed and A. Davis for photography. This work was supported by the Gatsby Charitable Foundation, the European Union Sixth Framework Programme Integrated Project SIROCCO (grant LSHG-CT-2006-037900), a postdoctoral fellowship from the National Science Foundation (to R.A.M.), and Commonwealth and NSERC Scholarships (to C.W.M.). D.C.B. is funded as a Royal Society Research Professor.

Author Contributions Experiments were designed by R.A.M. and D.C.B. and executed by R.A.M. with assistance from C.W.M. Bioinformatic and statistical analyses were performed by K.A.K., R.M.D. and D.J.S. The manuscript was prepared by R.A.M. and D.C.B.

Author Information Deep sequencing libraries are deposited in NCBI GEO (GSE15348). Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.C.B. (dcb40@cam.ac.uk).

METHODS

Plant growth. Plants were grown in controlled environment chambers under standard conditions with 16 h light per day. Flowers were manually inspected for anthesis to determine time of pollination/fertilization. For crosses, immature flowers were emasculated 24 h before manual pollination.

Northern blotting and siRNA quantification. Small (less than 200 nt) RNA was purified from more than 50 μg of total RNA with a mirVana miRNA isolation kit (Ambion) in accordance with the manufacturer's instructions. The sRNA fraction was then run on a 15% 19:1 acrylamide:bisacrylamide gel with 1 × TBE and 7 M urea and transferred to Hybond N^+ membrane (GE/Amersham) in 1 × TBE at 200 mA for 1 h. Membranes were crosslinked with 120 mJ ultraviolet before pre-hybridization in UltraHyb Oligo buffer (Ambion) at 35 °C for more than 15 min. Oligonucleotides (20 pmol) were labelled with [γ-^{32}P]ATP and T4 polynucleotide kinase before clean-up over a G-25 spin column (GE/Amersham). Hybridization of labelled oligonucleotides occurred in UltraHyb Oligo buffer at 35 °C, 6 h to overnight. Membranes were washed twice in 2 × SSC, 0.1% SDS at 35 °C before exposure to phosphor-storage screens. All northern blots were verified with independent biological replicates. Probe sequences are listed in Supplementary Table 2.

Three biological replicates each of wild-type, *nrpd1a-4*/*NRPD1A* and *NRPD1A*/*nrpd1a-4* were quantified blind with standard software. p4-siRNA or miR166 signals were normalized to $tRNA^{met}$ signals as a loading control. Values for locus 00687, locus 08002, locus 02815 and cluster 55 were combined to generate a 'p4-siRNA' value.

Statistical methods. In a three-way analysis of variance of p4-siRNA accumulation measured by northern blot quantification, differences in the factor of interest, 'heterozygote', were very highly significant ($F_{2,48} = 16.28$; $P < 10^{-4}$). The factors were 'tissue' (closed flowers, siliques five days after anthesis, fixed effects), 'locus' (four sRNA loci, random effects) and 'genotype' (*nrpd1a-4*/*NRPD1A*, *NRPD1A*/*nrpd1a-4* heterozygotes, wild type, fixed effect, fixed effect); there were three biological replicates (see Fig. 2b and Supplementary Fig. 2). Two orthogonal contrasts were made. The wild type was very highly significantly different from the two reciprocal heterozygotes ($F_{1,48} = 26.1$; $P < 10^{-6}$) and the reciprocal heterozygotes were different from one another ($F_{1,48} = 6.4$; $P < 0.05$). The aov function of R version 2.7.1 (2008-06-23) was used.

RT–PCR CAPS analysis. Total RNA from crossed siliques five days after fertilization was treated with DNase for 1 h at 37 °C before DNase inactivation (Ambion DNase-free kit). RNA (1 μg) was reverse transcribed with SuperScript III Reverse transcriptase (Invitrogen) in accordance with the manufacturer's instructions. Reverse transcription reactions and no enzyme controls were diluted 1:20 before PCR amplification with the primers listed in Supplementary Table 3. In some instances, additional primers were used for a preliminary PCR which was diluted and used as template for a nested PCR. Reactions were digested for 3 h to overnight with the noted enzyme and resolved on 2% agarose or 6% polyacrylamide gels as appropriate. Similar results were obtained in independent biological replicates. *FWA* RT–PCRs were performed as described[17].

Illumina sequencing of sRNA. Total RNA (30 μg) from crossed siliques five days after fertilization was run over a mirVana miRNA isolation column to enrich for RNAs more than 200 nt in length. 6.5 pmol of 5′ adaptor (Illumina small RNA sample prep kit) was ligated to this sRNA fraction in the following 10-μl reaction: RNA, adaptor, 1 × T4 RNA ligase buffer, 10 U of T4 RNA ligase (New England Biolabs), 40 U of RNaseOut (Invitrogen); 37 °C, 1 h. The reaction was run on a 15% 19:1 acrylamide:bisacrylamide gel with 7 M urea and 0.5 × TBE, and a band from 40 to 60 nt was excised. RNA was eluted from the gel fragment overnight in 1 × NEBuffer 2 (New England Biolabs). After precipitation, a second ligation was performed as above with 6 pmol of 3′ adaptor (Illumina small RNA sample prep kit), and run on a similar polyacrylamide gel. RNA was eluted as above from a gel slice of 60–80 nt and precipitated. RNA was reverse-transcribed in the following 10-μl reaction: adaptor-ligated RNA, 50 pmol of Small RNA RT-primer (Illumina small RNA sample prep kit), 200 U of SuperScript III RT (Invitrogen), 1 × first-strand buffer, 625 μM dNTPs, 10 mM dithiothreitol, 20 U of RNaseOut (Invitrogen); 50 °C, 1 h. A 4-μl sample of this reaction was PCR-amplified in a 200-μl reaction: 1 × Phusion High Fidelity PCR buffer, 4 U of Phusion Hot Start High Fidelity DNA Polymerase (Finnzymes), 25 pmol of each Small RNA PCR primer 1 and 2 (Illumina small RNA sample prep kit) and 250 μM dNTPs. Cycling conditions were 98 °C for 30 s, followed by 20 cycles of 98 °C for 10 s, 58 °C for 30 s and 72 °C for 20 s, and a final extension of 72 °C for 5 min. PCR products were precipitated and run on a 6% 19:1 acrylimide:bisacrylimide gel with 0.5 × TBE, and the appropriate band was excised and eluted.

Purified sRNA libraries were run on an Illumina Genome Analyser and the resulting reads were analysed. Reads were parsed, using Perl scripts, for the presence of a 3′ adaptor sequence and the ability to perfectly match either the Columbia or Landsberg *erecta* genome sequences.

***AtSN1* methylation.** Genomic DNA from leaf tissue of wild-type, *nrpd1a-4*, *nrpd1a-4*/*NRPD1A* and *NRPD1A*/*nrpd1a-4* plants was digested with 20 U of *Hae*III (1.5 μg in 30-μl reaction) for 2 h at 37 °C. The enzyme was inactivated for 20 min at 80 °C before further analysis. Digested DNA (150 ng) was amplified with the following primers: *Actin*, 5′-GCCATCCAAGCTGTTCTCTC-3′ and 5′-CCCTCGTAGATTGGCACAGT-3′; *AtSN1*, 5′-AAAATAAGTGGTGGTTG TACAAGC-3′ and 5′-CAACACATAAATGGCAATATACGATCC-3′. Three independent biological replicates were conducted.

LETTERS

A histone H3 lysine 36 trimethyltransferase links Nkx2-5 to Wolf–Hirschhorn syndrome

Keisuke Nimura[1], Kiyoe Ura[1], Hidetaka Shiratori[2], Masato Ikawa[3], Masaru Okabe[3], Robert J. Schwartz[4] & Yasufumi Kaneda[1]

Diverse histone modifications are catalysed and recognized by various specific proteins, establishing unique modification patterns that act as transcription signals[1,2]. In particular, histone H3 trimethylation at lysine 36 (H3K36me3) is associated with actively transcribed regions and has been proposed to provide landmarks for continuing transcription[3,4]; however, the control mechanisms and functions of H3K36me3 in higher eukaryotes are unknown. Here we show that the H3K36me3-specific histone methyltransferase (HMTase) Wolf–Hirschhorn syndrome candidate 1 (WHSC1, also known as NSD2 or MMSET) functions in transcriptional regulation together with developmental transcription factors whose defects overlap with the human disease Wolf–Hirschhorn syndrome (WHS)[5,6]. We found that mouse Whsc1, one of five putative Set2 homologues[2,7,8], governed H3K36me3 along euchromatin by associating with the cell-type-specific transcription factors Sall1, Sall4 and Nanog in embryonic stem cells, and Nkx2-5 in embryonic hearts, regulating the expression of their target genes. *Whsc1*-deficient mice showed growth retardation and various WHS-like midline defects, including congenital cardiovascular anomalies. The effects of *Whsc1* haploinsufficiency were increased in *Nkx2-5* heterozygous mutant hearts, indicating their functional link. We propose that WHSC1 functions together with developmental transcription factors to prevent the inappropriate transcription that can lead to various pathophysiologies.

Deletions of human chromosome 4p16.3 cause the dominant disorder WHS, which is characterized by cranio-facial malformations, learning disability, growth delays, heart defects and a diverse array of associated problems, many of which can be characterized as midline defects[5,6]. The considerable variability of the disorder, both genotypically and phenotypically, has led to the implication of multiple genes in the pathogenesis of WHS[6]. WHSC1, a protein encoded by one of several genes in the identified WHS critical region, is deleted in every known case of WHS and is dysregulated by t(4;14) translocations in lymphoid multiple myeloma[6,9]. The WHSC1 protein contains AWS–SET–PostSET domains that are highly conserved with yeast H3K36-specific methyltransferase Set2 (ref. 8) (Fig. 1a and Supplementary Fig. 1). However, the role of WHSC1 in chromatin function and its subsequent pathogenicity remain unclear.

To investigate whether WHSC1 possesses intrinsic HMTase activity, we performed *in vitro* HMTase activity assays with recombinant mouse Whsc1 (Supplementary Fig. 2a). Whsc1 preferentially methylated nucleosomal histone H3, and incorporation of histone H1 into the oligonucleosome inhibited histone methylation by Whsc1 (Fig. 1b). We determined the site specificity of Whsc1 with methylation-specific histone H3 antibodies. Whsc1 catalysed H3K36 monomethylation, dimethylation and trimethylation *in vitro* (Fig. 1c). Although amino-terminally deleted Whsc1 proteins have been shown to methylate histones H4K20 and H3K27 (ref. 10), full-length Whsc1 did not efficiently methylate nucleosomal histones at these sites *in vitro*.

To evaluate the HMTase activity of Whsc1 in the nucleus, the *Whsc1* locus was inactivated by deletion of its carboxy-terminal region, including the catalytic SET domain, in embryonic stem (ES) cells (Supplementary Fig. 3a, b). Northern blot and immunoblot analyses confirmed that no functional Whsc1 protein was expressed in homozygous *Whsc1* mutant ES cells (Supplementary Fig. 3c, d). *Whsc1*$^{-/-}$ ES cells retained an undifferentiated ES-cell morphology and normal cell growth. As shown in Fig. 1d, the absence of *Whsc1* did not significantly change H3K4 or H3K9 methylation. In contrast, the presence of H3K36me3, but not that of H3K36me2 or H3K36me1, was specifically decreased in *Whsc1*$^{-/-}$ ES cells. The level of H3K36me3 in *Whsc1*$^{-/-}$ ES cells was recovered when wild-type *Whsc1*, but not point-mutated (H1143G) inactive *Whsc1* (Supplementary Figs 1 and 2b), was stably expressed in these cells (Fig. 1d).

Endogenous Whsc1 localized in the ES cell nuclei, forming several small foci that did not overlap with 4,6-diamidino-2-phenylindole (DAPI)-stained heterochromatic foci, similar to the focal staining of H3K36me3 (Fig. 1e). We found that Whsc1-containing chromatin was specifically enriched in H3K36me3, but not in H3K36me2, in comparison with the total amount of histone H3 (Fig. 1f). These results show that Whsc1 is the major HMTase to regulate histone H3K36 trimethylation selectively at euchromatic regions in ES cells.

To uncover the function of Whsc1, we immunoaffinity-purified Whsc1-associated proteins from ES cells stably expressing Whsc1 with a C-terminal TAP-epitope tag (Whsc1–TAP). We identified the proteins in the major bands of SDS–PAGE analysis by mass spectrometry (Fig. 2a). Immunoblotting analyses confirmed that Whsc1 associated with Sall1 (ref. 11), a member of the zinc-finger transcription factor spalt (Sal)-like protein family, *O*-linked *N*-acetylglucosamine transferase (OGT)[12], and Brg1, an ATPase subunit of chromatin-remodelling complexes (Fig. 2b).

Sall1 interacts functionally with Sall4 (ref. 13), which exists as a complex with Nanog and histone deacetylase 1/2 (HDAC1/2) in ES cells[14,15]. Indeed, Sall4, Nanog and HDAC1, but not the ATPase subunit of the NuRD complex Mi-2, were identified as components of the Whsc1-associated proteins (Fig. 2b). We also observed a weak but significant interaction of Whsc1 with Brg1 and RNA polymerase II that was consistent with its euchromatic localization (Fig. 2b). Specific associations between Whsc1 and Sall1, Sall4 and Nanog were confirmed by co-immunoprecipitation assays (Supplementary Figs 4 and 5).

As shown in Fig. 2c, chromatin immunoprecipitation (ChIP) experiments revealed that Whsc1 localized to sites to which both

[1]Division of Gene Therapy Science, Osaka University Graduate School of Medicine, 2-2 Yamada-oka, Suita, Osaka 565-0871, Japan. [2]Developmental Genetics Group, Graduate School of Frontier Biosciences, and [3]Genome Information Research Center, Osaka University, 1-3 Yamada-oka, Suita, Osaka 565-0871, Japan. [4]Institute of Bioscience and Technology, Center for Molecular Development and Disease, 2121 West Holcombe Boulevard, Houston, Texas 77030, USA.

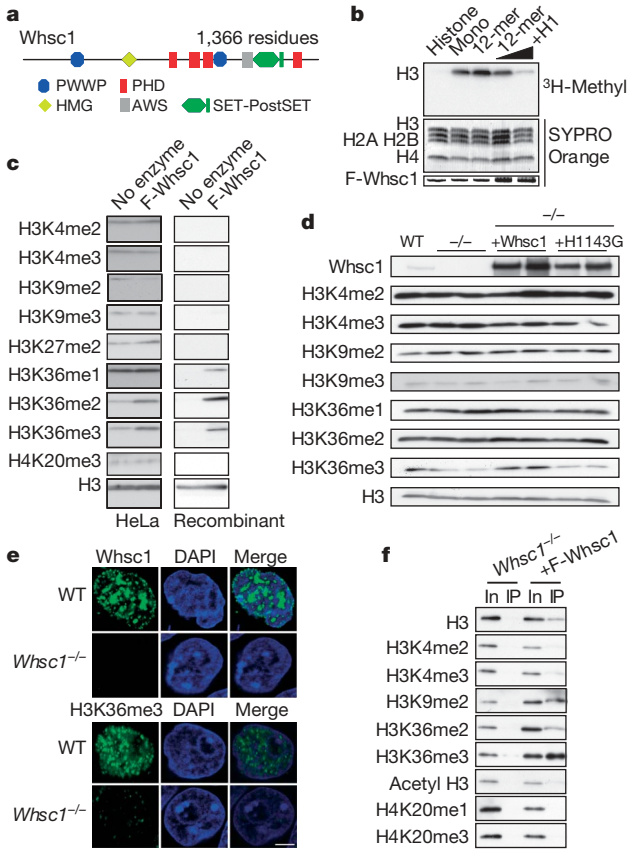


Figure 1 | Whsc1 methylates histone H3 on lysine 36. **a**, Schematic representation of mouse Whsc1. PWWP, domain with the conserved PWWP motif; HMG, high-mobility group box; PHD, plant homeotic domain; AWS, associated with SET; SET–PostSET, Su(Var)3-9, Enhancer-of-zeste, trithorax domain. **b**, HMTase activity of Flag–Whsc1 (F-Whsc1) towards free histone octamers, mononucleosomes (Mono), 12-mer nucleosome arrays and histone-H1-containing chromatin (12-mer+H1). Top panel, fluorogram; middle and bottom panels, SYPRO Orange staining of proteins. **c**, Whsc1 catalyses H3K36 monomethylation, dimethylation and trimethylation *in vitro*. HeLa nucleosomes or oligonucleosomes reconstituted from recombinant core histones were used as substrates for HMTase assays and analysed by immunoblotting with the indicated specific histone H3 antibodies. **d**, Methylation status of histone H3 in wild-type (WT) or *Whsc1*$^{-/-}$ ES cells and in *Whsc1*$^{-/-}$ ES cells stably expressing either Whsc1 or mutant Whsc1 (H1143G). Whole cell extracts were analysed by immunoblotting and the blots were probed with the antibodies indicated. **e**, Localization of Whsc1 and H3K36me3 in wild-type and *Whsc1*$^{-/-}$ ES cells. Cells were immunostained with anti-Whsc1 (chicken IgM) or with anti-H3K36me3. The cells were counterstained with DAPI to reveal chromatin DNA. Scale bar, 5 μm. **f**, Methylation status of Whsc1-containing chromatin. Whsc1-containing chromatin was purified from the nuclear extracts of *Whsc1*$^{-/-}$ ES cells stably expressing Flag–Whsc1 (Supplementary Fig. 4b). Immunoprecipitates (IP) and aliquots (0.7%) of the input protein (In) were analysed by immunoblotting with the antibodies indicated.

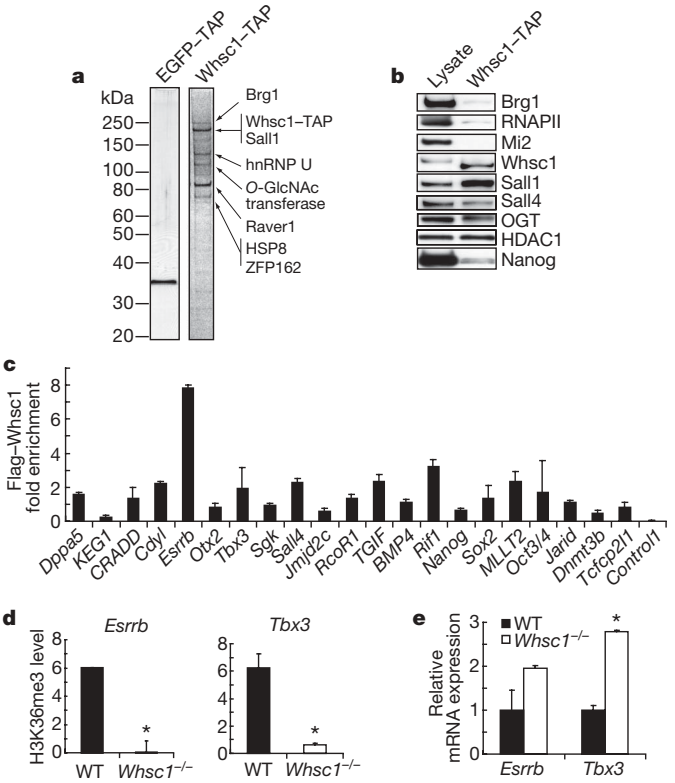


Figure 2 | Whsc1 associates with transcription factors to repress abnormal transcription. **a**, Silver staining of Whsc1–TAP-associated proteins purified from ES cells stably expressing Whsc1–TAP. Whsc1–TAP-associated proteins as identified by mass spectrometry are indicated at the right. ES cells expressing enhanced green fluorescent protein (EGFP)–TAP were used as a negative control. **b**, Purified Whsc1–TAP-associated proteins were analysed by immunoblotting with the antibodies indicated. **c**, Whsc1 occupancy at known Nanog–Sall4 co-binding loci and at one negative control region[14] was analysed by native ChIP assays. Flag–Whsc1 was immunoprecipitated from the nuclear extracts of *Whsc1*$^{-/-}$ ES cells stably expressing Flag–Whsc1 by using anti-Flag. Immunopurified DNA was quantified with real-time PCR (qPCR), and 0.7% of the digested chromatin was used as an input control. Data are from three independent experiments, and error bars indicate s.d. **d**, H3K36me3 occupancy on the *Esrrb* or *Tbx3* gene in wild-type (WT) and *Whsc1*$^{-/-}$ ES cells was analysed by ChIP experiments. Results were normalized to histone H3 and are shown as fold enrichment compared with the control1 value. Error bars indicate s.d. ($n = 3$). *$P < 0.01$. **e**, Quantitative RT–PCR analysis of *Esrrb* and *Tbx3* pre-mRNA in the nucleus. Reverse transcription of isolated nuclear RNA was performed with random primers. The *Esrrb* and *Tbx3* transcripts were analysed by qPCR. Values were normalized to the *Nanog*-independent gene *TRP53* value, and the wild type was set to 1. Error bars indicate s.d. ($n = 3$). *$P < 0.01$.

Nanog–Sall4 bind, but was not detectable at the Nanog–Sall4 non-binding control site (control 1)[14]. Two Nanog–Sall4 co-binding sites were located within the coding regions of the oestrogen-related receptor *Esrrb* and the T-box transcription factor *Tbx3*. The accumulation of H3K36me3 was significantly decreased around these Nanog–Sall4 co-binding sites in *Whsc1*$^{-/-}$ ES cells (Fig. 2c, d). We further analysed nuclear RNA and found that the levels of nuclear transcripts from these regions at *Esrrb* and *Tbx3* were increased on deletion of *Whsc1* (Fig. 2e). Because the Whsc1 complex contains HDAC1 (Fig. 2b), these results suggest that Whsc1 represses inappropriate transcription mediated by histone deacetylation, as is also seen in the yeast Set2 complex[16,17]. The distinctive feature of Whsc1 is its association with developmental transcription factors, whereas Set2 directly associates with the elongation form of RNA polymerase II through the Set2–Rpb1 interaction domain[18].

To elucidate the developmental and pathological role of Whsc1, we generated *Whsc1*-deficient mice (Supplementary Fig. 3e). Genotyping the offspring produced by interbreeding *Whsc1*$^{+/-}$ mice revealed significantly lower numbers of *Whsc1*$^{-/-}$ and *Whsc1*$^{+/-}$ mice than their expected Mendelian ratios at birth (Fig. 3a). We found that both homozygotes and heterozygotes were present at the expected frequencies and seemed morphologically normal at embryonic day (E)18.5 (Fig. 3a). However, E18.5 homozygous embryos were significantly smaller than their heterozygous or wild-type littermates (Supplementary Fig. 6). Furthermore, some homozygous E18.5 embryos did not breathe air after their extraction by caesarian section, suggesting the presence of developmental defects in these embryos.

After birth, *Whsc1*$^{-/-}$ mice showed growth retardation and died within 10 days (Fig. 3a, b). Most *Whsc1* heterozygotes survived and

a

Stage	WT	*Whsc1*$^{+/-}$	*Whsc1*$^{-/-}$
E15.5	16	33	17
E18.5	27	52	29
P0	49	46	4
P10	49	46	0

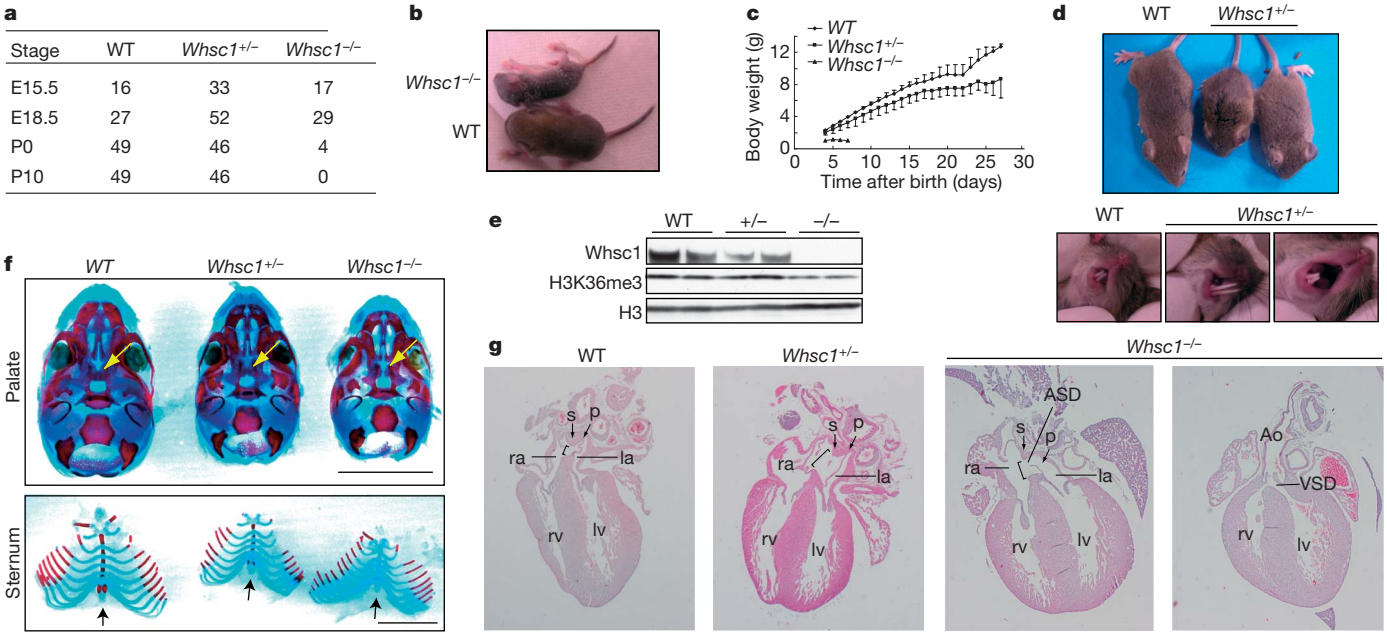


Figure 3 | The *Whsc1* gene is required for normal mouse development. **a**, Genotype analysis of embryos and newborn mice from *Whsc1*$^{+/-}$ intercrosses. **b**, Representative *Whsc1*$^{-/-}$ and wild-type (WT) littermates. One-week-old mice are shown. **c**, Body weights of representative *Whsc1*$^{-/-}$, *Whsc1*$^{+/-}$ ($n = 3$) and wild-type ($n = 2$) littermates after birth. Error bars represent s.d. **d**, Phenotypes displayed by *Whsc1*$^{+/-}$ mice. One one-month-old wild-type and two *Whsc1*$^{+/-}$ littermates are shown. Some *Whsc1* heterozygotes had severe growth defects and malocclusion incisors (middle). **e**, Expression of the Whsc1 protein and H3K36me3 levels were analysed by immunoblotting of nuclear extracts from two of each genotype of E15.5 embryos. The blots were probed with the antibodies indicated. **f**, Skeletal preparation of E18.5 embryos. Cartilage was stained with alcian blue, and ossified bone with alizarin red. Representative *Whsc1*$^{-/-}$, *Whsc1*$^{+/-}$ and wild-type littermates are shown. Yellow arrows, palates; black arrows, sternum. Scale bar, 5 mm. **g**, Histological analysis of *Whsc1* mutant embryonic hearts. Frontal sections from E18.5 embryos were stained with haematoxylin and eosin. Atrial septal defects (ASD) ($n = 7$ out of 7) and ventricular septal defects (VSD) ($n = 4$ out of 7) were observed in E18.5 *Whsc1*$^{-/-}$ embryos. The foramen ovales of *Whsc1*$^{+/-}$ embryos were larger than those of wild-type embryos (bracket). Hypoplasia of the septum secundum was observed in *Whsc1*$^{+/-}$ embryos ($n = 3$ out of 10). Ao, aortic root; lv, left ventricle; rv, right ventricle; la, left atrium; ra, right atrium; p, septum primum; s, septum secundum.

were fertile. However, the growth rates of *Whsc1*$^{+/-}$ mice were highly variable, and some of these mice showed severe growth and craniofacial defects similar to those seen in WHS[6] (Fig. 3c, d). It should be noted that the level of the Whsc1 protein in *Whsc1*$^{+/-}$ mice was only half of that in wild-type mice (Fig. 3e, top). These results indicate that the haploinsufficiency of the *Whsc1* gene in mice causes defects as a part of WHS.

Mouse *Whsc1* expression increased after E11 and was detected throughout the body at various intensities[9] (Supplementary Fig. 7). We found that H3K36 trimethylation was decreased in *Whsc1*$^{-/-}$ mice at E15.5, in comparison with wild-type mice (Fig. 3e). To investigate developmental defects, we stained the embryonic cartilage with alcian blue and the ossified bone with alizarin red, and found several deficiencies in midline fusion in mutant E18.5 embryos. In normal mice, the sternums fuse by E17.5 and subsequent ossification centres typically arise[19]. *Whsc1*$^{-/-}$ mice did not show any ossification centres at E18.5 (Fig. 3f, bottom). Even in *Whsc1*$^{+/-}$ mice, the appearance of ossification centres was markedly delayed. We further found incidences of cleft palate in *Whsc1*$^{-/-}$ mice, as is also seen in WHS[6] (Fig. 3f, top).

Patients with WHS often have congenital heart defects, including atrial and ventricular septal defects (ASD/VSD)[6]. We therefore next analysed cardiovascular development in *Whsc1* mutant embryos. Whsc1 was expressed throughout the embryonic heart, excluding the endocardial cushion (Supplementary Fig. 8). In wild-type mice at E18.5, the septum primum had grown to reach the atrial endocardium. We found that all *Whsc1*$^{-/-}$ hearts showed an ASD, and half also had a membranous VSD (Fig. 3g). Hypoplasia of the septum secundum was observed more frequently than in wild-type mice, even in heterozygous mutant mice (Fig. 3g and Supplementary Fig. 9). These results suggest that the loss of *Whsc1* causes a wide variety of midline defects, including the heart lesions seen in WHS, and leads to death after birth.

Congenital heart defects are caused by mutations in numerous genes, most of which encode cardiogenesis-specific transcription factors[20]. To examine whether a specific molecular interaction exists between Whsc1 and the cardiac transcription factors, we prepared the nuclear extracts from embryonic hearts at E12.5 and performed co-immunoprecipitation assays with an anti-Whsc1 antibody. Immunoblot analysis revealed a physical interaction between Whsc1 and a central transcriptional regulator of cardiac development, Nkx2-5 (ref. 21) (Fig. 4a and Supplementary Fig. 10). We detected a weak interaction of Whsc1 with Sall1, but not with Sall4, in embryonic hearts at E12.5 (Figs 2b and 4a).

The homeodomain factor *Nkx2-5* gene is currently thought to be one of the most commonly mutated genes in congenital heart diseases, including ASD and VSD[20,21]. In *Nkx2-5* mutant mouse hearts, various gene expressions are altered, both positively and negatively[21]. We found that genes known to be upregulated in *Nkx2-5* mutants, including *Pdgfra*[21], were significantly upregulated in *Whsc1*$^{-/-}$ mutant hearts (Fig. 4b and Supplementary Fig. 11). It should be noted that the expression of *Nkx2-5* or *Nppa*, known to be downregulated in *Nkx2-5* mutants[21], was not changed in *Whsc1* mutant hearts (Fig. 4b). We therefore speculated that Whsc1 collaborates with Nkx2-5 to repress their target genes directly. To examine this hypothesis, we performed ChIP assays at the *Pdgfra* locus (Fig. 4c) using E12.5 hearts. We found that the Nkx2-5 protein was highly enriched at the first exon of *Pdgfra*, in comparison with the reported Nkx2-5-binding site at the *Nppa* promoter[22] (Fig. 4d). Whsc1 also accumulated at the *Pdgfra* locus but was not detectable at *Nppa* (Fig. 4e). Consistent with the distribution of the Whsc1 protein, accumulation of H3K36me3 at the first exon of *Pdgfra* was decreased

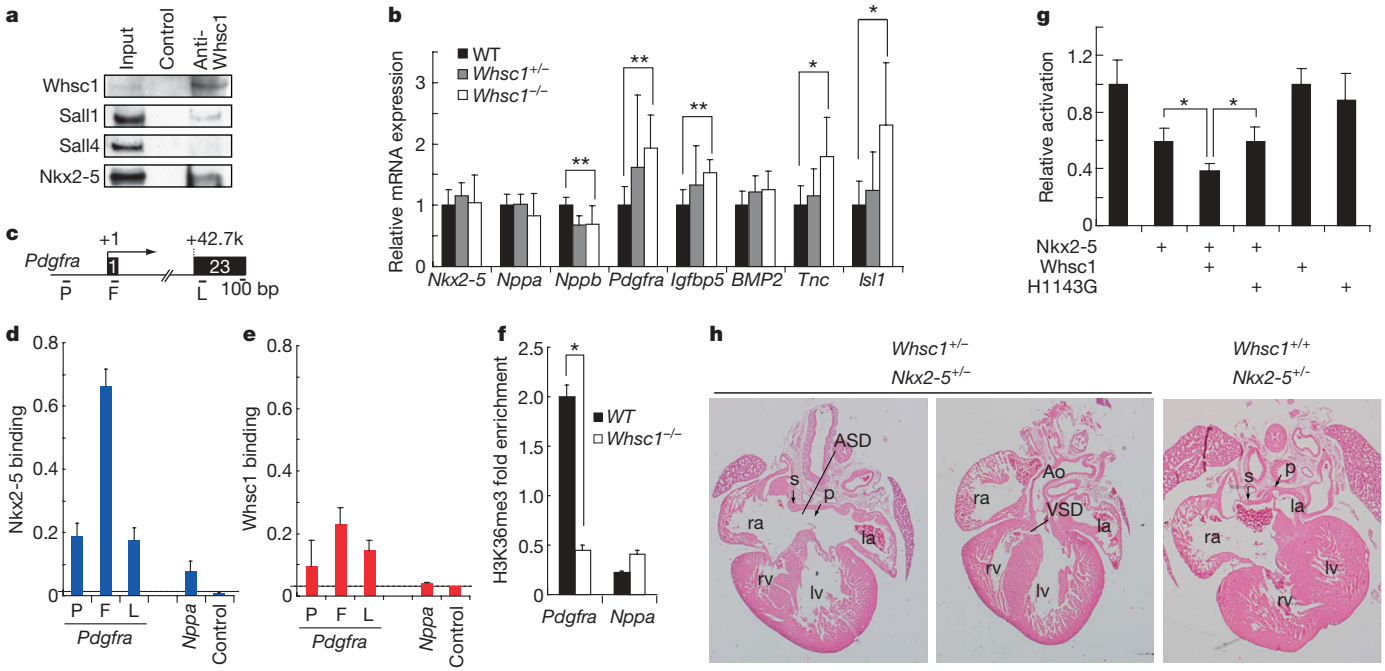


Figure 4 | Whsc1 is required for the appropriate transcription of *Nkx2-5*-dependent genes. **a**, Whsc1 was immunoprecipitated from the nuclear extracts of E12.5 hearts with an anti-Whsc1 or control antibody. Co-immunoprecipitates and aliquots (6%) of the input proteins were analysed by immunoblotting with the indicated antibodies. **b**, Quantitative RT–PCR analysis of the expressions of *Nkx2-5* and *Nkx2-5*-dependent genes[21] in the hearts of E12.5 embryos of the indicated genotypes. Values were normalized against expression of *Gapdh* (the gene encoding glyceraldehyde-3-phosphate dehydrogenase) and set to 1.0 for each gene in the wild type. Averages from six embryos of each genotype group are shown (Supplementary Fig. 11). Error bars represent s.d. $*P < 0.05$; $**P < 0.01$. **c**, Schematic representation of the *Pdgfra* gene. Locations of genomic regions analysed by ChIP assays are indicated. P, promoter; F, first exon; L, last exon 23. **d**, **e**, Native ChIP assays examining the occupancy of Nkx2-5 (**d**) and Whsc1 (**e**) at the *Pdgfra* regions indicated in **c**, the *Nppa* promoter, and the silent *histone H1foo* gene as a negative control in hearts from wild-type embryos at E12.5. Results are shown as the percentage enrichment relative to 0.5% input. Error bars represent s.d. ($n = 3$). Dashed line, background level. **f**, Native ChIP assays examining the occupancy of H3K36me3 at the first exon of *Pdgfra* and the promoter of *Nppa* in hearts from E12.5 embryos with the indicated genotypes. Results are normalized to histone H3. Error bars indicate s.d. ($n = 3$). $*P < 0.01$. **g**, Repressive effects of Whsc1 and Nkx2-5 on the *Pdgfra*-luciferase reporter construct in 10T1/2 cells. The reporter construct was co-transfected with the indicated expression constructs. Error bars represent s.d. ($n = 4$). $*P < 0.03$. **h**, Histological analysis of *Whsc1*$^{+/-}$ and *Nkx2-5*$^{+/-}$ embryonic hearts. Frontal sections from E18.5 embryos were stained with haematoxylin and eosin. Both ASDs and VSDs were observed in E18.5 hearts from *Whsc1*$^{+/-}$*Nkx2-5*$^{+/-}$ double heterozygous mutants ($n = 3$ out of 9), whereas no significant defects were observed in *Whsc1*$^{+/+}$*Nkx2-5*$^{+/-}$ embryonic hearts[24] ($n = 3$). Ao, aortic root; lv, left ventricle; rv, right ventricle; la, left atrium; ra, right atrium; p, septum primum; s, septum secundum.

in *Whsc1*$^{-/-}$ hearts (Fig. 4f). Co-localization of Whsc1 with Nkx2-5 was observed not only at *Pdgfra* but also at two other Nkx2-5-dependent gene loci, *Igfbp5* and *Isl1* (Fig. 4b and Supplementary Figs 12 and 13).

Transient reporter assays have shown that Nkx2-5 activates *Nppa* expression through its binding at the promoter[22]. We found that the activity of a luciferase reporter plasmid containing the *Pdgfra* promoter and the first exon (*Pdgfra*-luc) was repressed by Nkx2-5, and that Whsc1 significantly enhanced this repressive effect of Nkx2-5 on *Pdgfra* (Fig. 4g and Supplementary Fig. 14 and 15). In contrast, Whsc1 had little effect on the activity of the *Nppa* promoter (Supplementary 14b). Furthermore, a point mutation in the SET domain of Whsc1 (H1143G) abolished the repressive effect of Whsc1 on *Pdgfra*, despite the association of mutant Whsc1 with Nkx2-5 (Fig. 4g and Supplementary Fig. 16). Although we could not rule out the possibility that Whsc1 has functional non-histone substrates, these results suggest that Whsc1 is required for the proper functionality of Nkx2-5 to downregulate transcription of the target gene, presumably mediated by H3K36me3.

Finally, we crossed *Whsc1*$^{+/-}$ and *Nkx2-5*$^{+/-}$ mice[23] to investigate their functional link *in vivo*. Neither *Nkx2-5*$^{+/-}$ nor *Whsc1*$^{+/-}$ single-heterozygous mutant mice exhibited any significant defects in the septum primum or the interventricular septum of their hearts[24] at E18.5 (Figs 3g and 4h). In contrast, we found both an ASD and a VSD in one-third of embryonic hearts from *Whsc1*$^{+/-}$*Nkx2-5*$^{+/-}$ double-heterozygous mutants (Fig. 4h). Hypoplasia of the septum secundum was observed more frequently ($n = 5$ out of 9) and more severely in *Whsc1*$^{+/-}$*Nkx2-5*$^{+/-}$ than in single-heterozygous mutants. Thus, we have demonstrated here the genetic interaction between *Whsc1* and *Nkx2-5* during atrial/ventricular septal formation. Our findings partly explain the observation that the congenital heart malformations seen in both *Whsc1* mutant mice (Fig. 3g) and patients with WHS[6] are caused by dysfunction of Nkx2-5.

We have revealed a developmental and pathological link between H3K36 trimethyltransferase and transcription factors. Interactions of multiple transcription factors with Whsc1 could account for the variability of defects caused by the *Whsc1* deficiency. Our studies provide new insights into the molecular mechanism of Nkx2-5-dependent gene regulation in hearts, in which Whsc1 negatively modulates the transcriptional activity of Nkx2-5 (Fig. 4). Because Nkx2-5 regulates transcription by cooperating with various cardiac transcription factors[20,22,25], Whsc1 might function in the tuning of these transcriptional networks. The association of Whsc1 with HDAC, OGT[12] and the heterogeneous nuclear ribonucleoprotein Raver1 (Fig. 2a), which regulates the alternative splicing of particular genes[26], suggests that Whsc1 functions as more than a simple transcriptional repressor[10] during the stepwise process of transcription. Although further studies are needed to explain the molecular function of Whsc1, our findings point to a new direction for the understanding and treatment of dysregulated transcription, in which WHSC1 functions together with developmental transcription factors to prevent transcriptional pathophysiologies.

METHODS SUMMARY

ES cells and mice. *Whsc1*$^{-/-}$ ES cells were generated with J1 ES cells. Whsc1-deficient mice were generated with D3 ES cells in accordance with standard protocols, and backcrossed to C57BL/6 from two to eight times. The targeting strategy is shown in Supplementary Fig. 2a. Nkx2-5-deficient mice were generated as described[23].

Co-immunoprecipitation analysis. Nuclear extracts were prepared from the DNase I-treated nuclei of ES cells, embryos and embryonic hearts, and then immunoprecipitated as described[27].

Native ChIP assays. Isolated nuclei[27] were treated with 2,400 U ml^{-1} micrococcal nuclease in 250 μl of nuclear isolation buffer containing 200 mM NaCl at 25 °C for 30 min. SYBER Green PCR master mix (TaKaRa) was used for quantitative real-time PCR. Primer sequences are shown in Supplementary Table 1.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 November 2008; accepted 23 April 2009.
Published online 31 May 2009.

1. Sims, R. J. & Reinberg, D. Histone H3 Lys 4 methylation: caught in a bind? *Genes Dev.* **20,** 2779–2786 (2006).
2. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–719 (2007).
3. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
4. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
5. Hirschhorn, K. & Cooper, H. L. Chromosomal aberrations in human disease. A review of the status of cytogenetics in medicine. *Am. J. Med.* **31,** 442–470 (1961).
6. Bergemann, A. D., Cole, F. & Hirschhorn, K. The etiology of Wolf–Hirschhorn syndrome. *Trends Genet.* **21,** 188–195 (2005).
7. Strahl, B. D. *et al.* Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. *Mol. Cell. Biol.* **22,** 1298–1306 (2002).
8. Sun, X. J. *et al.* Identification and characterization of a novel human histone H3 lysine 36-specific methyltransferase. *J. Biol. Chem.* **280,** 35261–35271 (2005).
9. Stec, I. *et al.* WHSC1, a 90 kb SET domain-containing gene, expressed in early development and homologous to a *Drosophila* dysmorphy gene maps in the Wolf-Hirschhorn syndrome critical region and is fused to IgH in t(4;14) multiple myeloma. *Hum. Mol. Genet.* **7,** 1071–1082 (1998).
10. Marango, J. *et al.* The MMSET protein is a histone methyltransferase with characteristics of a transcriptional corepressor. *Blood* **111,** 3145–3154 (2008).
11. Kohlhase, J., Wischermann, A., Reichenbach, H., Froster, U. & Engel, W. Mutations in the SALL1 putative transcription factor gene cause Townes-Brocks syndrome. *Nature Genet.* **18,** 81–83 (1998).
12. Shafi, R. *et al.* The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proc. Natl Acad. Sci. USA* **97,** 5735–5739 (2000).
13. Sakaki-Yumoto, M. *et al.* The murine homolog of SALL4, a causative gene in Okihiro syndrome, is essential for embryonic stem cell proliferation, and cooperates with Sall1 in anorectal, heart, brain and kidney development. *Development* **133,** 3005–3013 (2006).
14. Wu, Q. *et al.* Sall4 interacts with Nanog and co-occupies Nanog genomic sites in embryonic stem cells. *J. Biol. Chem.* **281,** 24090–24094 (2006).
15. Liang, J. *et al.* Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. *Nature Cell Biol.* **10,** 731–739 (2008).
16. Carrozza, M. J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123,** 581–592 (2005).
17. Li, B. *et al.* Combined action of PHD and chromo domains directs the Rpd3S HDAC to transcribed chromatin. *Science* **316,** 1050–1054 (2007).
18. Li, M. *et al.* Solution structure of the Set2-Rpb1 interacting domain of human Set2 and its interaction with the hyperphosphorylated C-terminal domain of Rpb1. *Proc. Natl Acad. Sci. USA* **102,** 17636–17641 (2005).
19. Liu, K. J., Arron, J. R., Stankunas, K., Crabtree, G. R. & Longaker, M. T. Chemical rescue of cleft palate and midline defects in conditional GSK-3β mice. *Nature* **446,** 79–82 (2007).
20. Bruneau, B. G. The developmental genetics of congenital heart disease. *Nature* **451,** 943–948 (2008).
21. Prall, O. W. *et al.* An Nkx2-5/Bmp2/Smad1 negative feedback loop controls heart progenitor specification and proliferation. *Cell* **128,** 947–959 (2007).
22. Hiroi, Y. *et al.* Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation. *Nature Genet.* **28,** 276–280 (2001).
23. Moses, K. A., DeMayo, F., Braun, R. M., Reecy, J. L. & Schwartz, R. J. Embryonic expression of an Nkx2-5/Cre gene using ROSA26 reporter mice. *Genesis* **31,** 176–180 (2001).
24. Biben, C. *et al.* Cardiac septal and valvular dysmorphogenesis in mice heterozygous for mutations in the homeobox gene Nkx2-5. *Circ. Res.* **87,** 888–895 (2000).
25. Koshiba-Takeuchi, K. *et al.* Cooperative and antagonistic interactions between Sall4 and Tbx5 pattern the mouse limb and heart. *Nature Genet.* **38,** 175–183 (2006).
26. Gromak, N. *et al.* The PTB interacting protein raver1 regulates α-tropomyosin alternative splicing. *EMBO J.* **22,** 6356–6364 (2003).
27. Nimura, K. *et al.* Dnmt3a2 targets endogenous Dnmt3L to ES cell chromatin and induces regional DNA methylation. *Genes Cells* **11,** 1225–1237 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Richmond, R. Nishinakamura, H. Kurumizaka and M. Shirai for providing reagents and mice; S. Khochbin, S. Hirose, J. Godde and J. Takeuchi for their critical reading of the manuscript; and H. Niwa, H. Hamada, M. Yamamoto, S. Ohishi, H. Nakagami and members of the GTS laboratory for discussion and support. This work was supported by grants from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (MEXT) and the Naito Foundation, and by funds from Osaka University for female researchers.

Author Contributions Y.K. provided support and general guidance for this work. K.U. planned and organized the project. K.N. designed and performed experiments. H.S. performed histological analysis of hearts. R.J.S., M.I. and M.O. contributed mouse resources. K.U. wrote the paper with K.N. and Y.K.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.U. (kiyoeura@gts.med.osaka-u.ac.jp) or Y.K. (kaneday@gts.med.osaka-u.ac.jp).

METHODS

Whsc1 constructs. Whsc1 complementary DNA was amplified from ES-cell cDNA by PCR and then introduced into pENTR (Invitrogen) using Gateway technology to generate the plasmid pENTR/Whsc1. Deletions or point mutations were introduced into pENTR/Whsc1 by PCR (Supplementary Table 1). To express Flag–Whsc1 and mutant proteins in ES and 10T1/2 cells, coding regions were introduced into pCAGIP-flag-gw[28].

Histone methylation assays. Native core histones, nucleosomes and histone H1 were purified from HeLa cells[29], and recombinant human core histones were prepared as described previously[30]. Monomer and 12-mer nucleosomes were reconstituted onto the 1×208 base pairs (bp) and 12×208 bp of the '601' sequence[31,32] by salt dialysis. Histone H1 was assembled into 12-mer nucleosomes by using NAP-1 as a linker histone chaperone[33].

The full-length cDNAs of wild-type Whsc1 and point-mutated Whsc1-H1143G were introduced into the pFastBac-FlagC-gw plasmid produced from pFastBac1 (Invitrogen). The F-Whsc1 expression vectors were transformed into DH10Bac (Invitrogen) and expressed in Sf9 (*Spodoptera frugiperda*) cells with the use of the Bac-to-Bac Baculovirus Expression System (Invitrogen). Recombinant proteins were purified with an anti-Flag M2 affinity gel and 3XFLAG peptide (Sigma).

The HMTase assay was performed as described previously[34]. Histone substrates (400 ng each) were incubated in a 20-μl reaction buffer at 30 °C for 4 h. Samples were separated by 15–25% SDS–PAGE gradient electrophoresis, stained with SYPRO Orange (Molecular Probes) and exposed to X-ray film (Kodak) or analysed by western blot analysis with modification-specific antibodies.

ES-cell culture and transfection. Wild-type J1 and mutant ES cells were maintained without feeder cells[27]. Whsc1 proteins were stably expressed in ES cells by electroporation.

Antibodies. A rabbit anti-Whsc1 antibody for western blotting was purified from the serum of a rabbit immunized against the recombinant His-tagged C-terminal part of Whsc1 (residues 630–1366), which was expressed in *Escherichia coli* and purified in accordance with the manufacturer's instructions (Qiagen). A chicken IgM anti-Whsc1 antibody for immunofluorescence analysis was screened from a DT40 library by using recombinant Flag-tagged full-length Whsc1 as bait with the ADLib system (Chiome)[35]. The anti-Whsc1 antibody for immunoprecipitation was purified from the serum of a rabbit immunized against the peptides C+RETQESERKPPPYKH and C+SSTKTEKPFPESLKS. The other antibodies used were as follows: anti-histone H3, anti-H3K4me3, anti-H3K9me2, anti-H3K9-me3, anti-H3K36me1, anti-H3K36me3, anti-H4K20me1 and anti-H4K20me3 (Abcam), anti-H3Ac, anti-H3K4me2, anti-H3K27me2, anti-H3K36me2 and anti-HDAC1 (Upstate), anti-Brg1, anti-OGT, anti-Oct3/4, anti-RNAPII and anti-Nkx2-5 (Santa Cruz Biotechnology), anti-Nanog (Reprocell), anti-Sall1 and anti-Sall4 (gifts from R. Nishinakamura)[13] and anti-Mi2 (gift from P. Wade)[36].

Targeted disruption of the *Whsc1* gene. A targeting vector with a promoter trap and DT-A selection cassette was designed to replace exons 13–21 of the mouse *Whsc1* gene, which contains the SET domain, with an internal ribosomal entry site (IRES)-β-geo cassette with the use of Gateway technology (Invitrogen). The 11.7-kilobase (kb) fragment containing intron 9 to exon 12 was introduced into pDONR P4-P1R (Invitrogen), the 2.1-kb fragment containing intron 21 and exon 22 was introduced into pDONR P2R-P3, and the IRES-β-geo cassette was introduced into pENTR1A. These three vectors were introduced into pDEST R4-R3 containing a DT-A sequence by using the MultiSite Gateway Three-fragment Vector Kit (Invitrogen). Homozygous *Whsc1* mutant ES cell lines were obtained by the selection of heterozygous mutant cells in a medium containing a high concentration of G418.

Immunofluorescence analysis. ES cells were fixed with 4% paraformaldehyde and immunostained as described[27].

Skeletal preparation, and 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-Gal) and haematoxylin stainings. Alcian blue and alizarin red staining of bones and cartilage, X-Gal staining of embryos, and haematoxylin staining of embryonic hearts were performed in accordance with established protocols.

Tandem affinity purification and mass spectrometry. The cDNAs of *Whsc1* and *EGFP* were introduced into pCAGIP-gw-TAP with the use of Gateway technology (Invitrogen). Whsc1–TAP and EGFP–TAP complexes were purified from ES-cell nuclear extracts by using TAP technology, as described[37]. The purified proteins were precipitated with trichloroacetic acid and separated on SDS–PAGE gels. Protein bands stained with silver were excised from the gel, in-gel digested with trypsin and analysed by tandem liquid chromatography/mass spectroscopy (LC–MS/MS).

Native ChIP assays. Isolated nuclei[27] from 10^8 ES cells were treated at 25 °C for 30 min with 933.3 U ml^{-1} micrococcal nuclease in 7.5 ml of a nuclear isolation buffer containing 200 mM NaCl. SYBER Green PCR master mix (TaKaRa) was used for quantitative real-time PCR. Primer sequences are shown in Supplementary Table 1 or were as described previously[14].

Quantitative RT–PCR. Total RNA was extracted with Isogen (Wako). Reverse transcription was performed with SuperScript III (Invitrogen) and analysed as described[28]. TaqMan probes were purchased from ABI.

Luciferase assays. The *Pdgfra*-luciferase construct (−2000 to +491) was generated by PCR and cloned into pGL3-basic (Promega). The cDNAs of *Nkx2-5*, *Whsc1* and point-mutated *Whsc1* (H1143G) were cloned into pCAGIP-flag-gw. The *Pdgfra*-luciferase or *Nkx2-5*-luciferase reporter constructs[22,38] were co-transfected with the expression constructs into C3H 10T1/2 murine mesenchymal progenitor cells by FuGENE 6 (Roche) as described previously[28]. Luciferase activity was measured by the Dual Luciferase Reporter Assay system (Promega).

28. Hayashi, M. *et al.* Comparative roles of Twist-1 and Id1 in transcriptional regulation by BMP signaling. *J. Cell Sci.* **120,** 1350–1357 (2007).
29. Ura, K. & Kaneda, Y. Reconstitution of chromatin *in vitro*. *Methods Mol. Biol.* **181,** 309–325 (2001).
30. Tachiwana, H., Osakabe, A., Kimura, H. & Kurumizaka, H. Nucleosome formation with the testis-specific histone H3 variant, H3t, by human nucleosome assembly proteins *in vitro*. *Nucleic Acids Res.* **36,** 2208–2218 (2008).
31. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276,** 19–42 (1998).
32. Dorigo, B., Schalch, T., Bystricky, K. & Richmond, T. J. Chromatin fiber folding: requirement for the histone H4 N-terminal tail. *J. Mol. Biol.* **327,** 85–96 (2003).
33. Saeki, H. *et al.* Linker histone variants control chromatin dynamics during early embryogenesis. *Proc. Natl Acad. Sci. USA* **102,** 5697–5702 (2005).
34. Nishioka, K. & Reinberg, D. Methods and tips for the purification of human histone methyltransferases. *Methods* **31,** 49–58 (2003).
35. Seo, H. *et al.* Rapid generation of specific antibodies by enhanced homologous recombination. *Nature Biotechnol.* **23,** 731–735 (2005).
36. Wade, P. A. *et al.* Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nature Genet.* **23,** 62–66 (1999).
37. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17,** 1030–1032 (1999).
38. Bruneau, B. G. *et al.* A murine model of Holt–Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell* **106,** 709–721 (2001).

ADDENDUM

doi:10.1038/nature08164

Ionic high-pressure form of elemental boron

Artem R. Oganov, Jiuhua Chen, Carlo Gatti, Yanzhang Ma, Yanming Ma, Colin W. Glass, Zhenxian Liu, Tony Yu, Oleksandr O. Kurakevych & Vladimir L. Solozhenko

Nature **457**, 863–867 (2009)

This Letter presents the results of high-pressure experiments and *ab initio* evolutionary crystal structure predictions, and found a new boron phase that we named γ-B_{28}. This phase is comprised of icosahedral B_{12} clusters and B_2 pairs in a NaCl-type arrangement, stable between 19 and 89 GPa, and exhibits evidence for charge transfer (for which our best estimate is $\delta \approx 0.48$) between the constituent clusters to give $(B_2)^{\delta+}(B_{12})^{\delta-}$. We have recently found that the same high-pressure boron phase may have given rise to the Bragg reflections reported by Wentorf in 1965 (ref. 1), although the chemical composition was not analysed and the data (subsequently deleted from the Powder Diffraction File database) seems to not have been used to propose a structure model. We also note that although we used the terms 'partially ionic' and 'ionic' to emphasize the polar nature of the high-pressure boron phase and the influence this polarity has on several physical properties of the elemental phase, the chemical bonding in γ-B_{28} is predominantly covalent.

We acknowledge N. Dubrovinskaia, L. Dubrovinsky, E. Yu Zarechnaya, Y. Filinchuk, D. Chernyshov, V. Dmitriev, A. S. Mikhaylushkin, I. A. Abrikosov & S. I. Simak for drawing these issues to our attention.

1. Wentorf, R. H. Boron: another form. *Science* **147**, 49–50 (1965).

ERRATUM

doi:10.1038/nature08202

Slow earthquakes triggered by typhoons

ChiChing Liu, Alan T. Linde & I. Selwyn Sacks

Nature 459, 833–836 (2009)

In this Letter, address 1 was incorrectly listed. This oversight has now been rectified.

PROSPECTS

Debating the literature

A new approach can revitalize literature updates, building skills and teamwork at the same time. B. Harihara Venkatraman, Dipankar Basak and Dhandapani Venkataraman report.

Research-group meetings, presentations, journal clubs, online blogs and literature seminars — these are the usual ways for labs to keep abreast of the scientific literature. But they often lack in-depth contributions from many participants. Postdocs and students may view literature updates as a chore rather than as a stimulating intellectual exercise.

To ensure incisive and deliberative discussions among all participants, we have pursued a new approach — adopting debates as the format for literature updates during our weekly group meetings. These debates provide not only an opportunity to learn about the topical literature, but also a way for group members to develop vital soft skills, such as working as a team and presenting data.

Our research group, undergraduates included, is divided into three sub-groups: 'pro', 'con' and 'jury'. The members switch sub-groups each week. The jury chooses a research paper published as a communication in the most recent edition of *Journal of the American Chemical Society* (*JACS*), a week before the group meeting. The pro group argues that the paper meets the publication criteria in terms of novelty, significance, broad interest and scholarly presentation. The con group argues otherwise. Both teams base their arguments on a critical analysis of the data provided in the paper and prior literature.

Each team has three minutes per speaker and three minutes for closing arguments. After closing arguments, the jury can ask questions related either to the arguments or to the paper. Each group has two minutes to answer the jury's question and to defend against the arguments made by the other group. Then jury members vote for a winner in a secret ballot. The principal investigator does not participate but offers post-debate comments on the validity of arguments made, essential points that should have been discussed and areas that need improvement.

Although each paper has already been published in *JACS*, the jury has historically shown no bias for the pro group. The debates have actually led us to dispute the conclusions of a few papers, including the accuracy of cartoons or figures in the papers. In a couple of cases, we contacted the principal authors to provide our post-debate viewpoints.

Through such debates, our group members have developed the ability to work as a team, dissect scientific arguments, listen to opposing viewpoints without prejudgements, overcome personal bias on a topic and provide a professional response based on the data. They have learned to engage in constructive criticism and adopt a healthy scepticism. These skills are valuable not only for their own research projects, but also in the development of students for their chosen careers. ■

B. Harihara Venkatraman and Dipankar Basak are graduate students in the department of chemistry at the University of Massachusetts Amherst. Dhandapani Venkataraman is their adviser.

POSTDOC JOURNAL

Expect the unexpected

Labs are often serene, orderly places. But disaster may lurk. Recently, I was at my desk, happily reading an article about how transcription remodelling factors regulate nucleosome dynamics, when I heard a commotion next door. I ran over to find that a pipe had burst. Water was spewing everywhere, flooding the lab within minutes.

Researchers abandoned their experiments as everyone worked together to staunch the inundation. Someone built a makeshift dam while others mopped up gallons and gallons of water, pouring it all down the sink. Finally, after one of us found and shut off the emergency valve, the deluge slowed and stopped. We all sighed with relief when we found that no equipment had been damaged.

That wasn't our only near-disaster. A few months ago, the −80 °C freezer in our lab broke, triggering a mad scramble to transfer precious antibodies, strains and reagents to a backup freezer. A lab's deep freezer often contains years of research. And then there's the terror of losing years of stored digital data to a hard-drive crash. Last year a friend of mine had a hard-drive malfunction, lost part of his thesis and had to start again from scratch.

What have I learned from these unfortunate situations? Be vigilant about protecting laboratory reagents and research data. And don't be lulled into a false sense of security by a quiet, tranquil lab bathed in the soft hum of a working freezer. Expect the unexpected. ■

Bryan Venters is a postdoctoral fellow at the Center for Eukaryotic Gene Regulation at Pennsylvania State University, University Park.

IN BRIEF

Boon for biotech

There's life yet for biotechnology firms, and 15 of them have special promise, according to FierceBiotech, an industry bulletin that covers the biotech sector. FierceBiotech named six firms from Greater Boston, Massachusetts; one from the United Kingdom; five from California and another three from elsewhere in the United States in its list of the year's top 15 biotechs (see www.fiercebiotech.com). The financial downturn made the cream of the crop easier to identify, FierceBiotech said, because only a select group of biotechs have thrived in the current climate. FierceBiotech based its selections on such criteria as the company's drug development experience, the novelty or efficacy of its product and the transparency of its management.

Student enrolment up

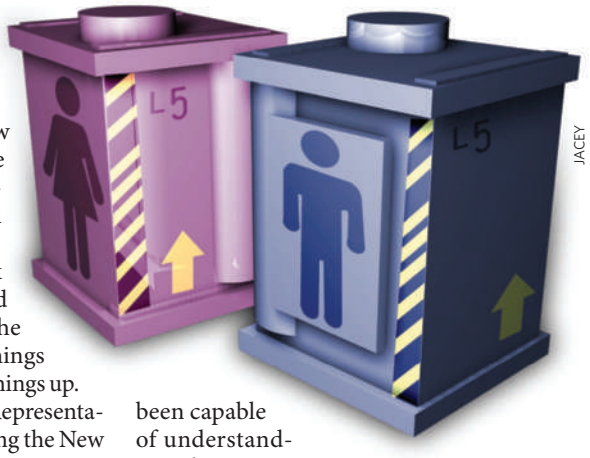
Graduate-level enrolment in science and engineering in the United States reached 516,199 in 2007, up 3.3% on 2006, according to the most recent survey from the National Science Foundation. The year-on-year increase was the highest since 2002 and nearly double that of the year before. Of these students, nearly three-quarters (72%) were enrolled full time and almost three quarters (71%) were US citizens or permanent residents. More than half (56%) were men. Among US citizens and permanent residents, men and women enrolled in almost equal numbers, but among temporary visa holders, two-thirds were men. New full-time enrolments of non-US students were up 8.3%.

Plain English

Two young researchers have launched a free manuscript-editing service to help scientists from developing countries get papers published in English-language journals. Justin Chakma, a research analyst at the McLaughlin-Rotman Centre for Global Health in Toronto, Canada, and Alexander Patananan, a biochemistry doctoral student at the University of California, Los Angeles, set up SciEdit earlier this year as a spin-off of the *Journal of Young Investigators*, an international student-run, peer-reviewed journal. In choosing which papers to improve, it assesses the merit of the science, not just the use of English. To date, SciEdit has accepted 10 manuscripts from India, South Africa, Pakistan, China and Nepal.

Goliath

It's all about timing.



JACEY

Bruce W. Ferguson

The Spacegard XP survey overlooked Goliath because there was nothing to fear until a dead comet caromed off a nickel-iron buckle in the asteroid belt. A 20-kilometre fragment tumbled crazily on its new trajectory.

Ashen-faced astronomers announced the dinosaur-killer would create a new ocean basin in the Sahel. Time to impact: 1,001 nights. By Cope's rule nothing on Earth larger than a gerbil would survive, and even the gerbils would have to draw lots.

Global response was swift. "We shall mobilize every resource to overcome this dire threat," declared the US president. "Immediately," agreed the UN secretary-general. "Dinosaurs are a Darwinist myth," added a watchdog group in Kent.

To ensure human survival, Goliath had to be diverted and Earth fortified. For the first time in Earth's history, the vast planetary resources of an advanced technological society united in a single mission other than winning the World Cup.

On Day 5, the UN announced an intercept plan. Aerospace contractors would cluster the world's ICBMs into larger rockets. Dozens of new launch sites would be built around the world. Hundreds of nuclear bombs would be launched. Stand-off detonation would spall Goliath's surface, blasting relativistic ions off one side of the monster to nudge it the other way.

Time was of the essence. Governments quickly awarded massive sole-source contracts. After a few months, the initial funding rush slowed as bills were introduced in Congress to increase union labour content. The European Union debated member-state budget allocations. Contracts were negotiated more carefully, then renegotiated. Machine parts sat for weeks on a West African dock until a quiet payment was made to the Premier's uncle.

The Martha's Vineyard Beautification Committee picketed the construction of rocket platforms off the Massachusetts coast, claiming the launch platforms were even uglier than windmills. By Day 200 approximately 30,612 environmental impact statements had been filed. Rocket construction slowed further.

Abrupt, overjoyed relief! New Zealand researchers calculated Goliath would miss Earth by a whisker. They proposed a follow-on grant to make absolutely sure. In Pasadena, Carnegie Observatories noted hesitantly that the New Zealand model reversed the sign of gravitational acceleration, thereby assuming Earth and Goliath would repel each other. Spinning a globe, a Fox News commentator pointed out that although gravity in the Northern Hemisphere held things down, in New Zealand it held things up. Galvanized, the US House of Representatives swiftly passed a bill funding the New Zealand proposal and imposing a 90% tax rate on Carnegie scientists.

Whispers grew that the intercept plan was fatally flawed. On Day 337 the world's greatest orbital dynamicist, a 50-year-old cigar smoker named Enrico who headed the UN's Intercept Task Force, scowled in fierce concentration as he walked slowly down Rome's Via Veneto. Suddenly his face brightened. "Eureka!" he exulted, striding directly into the path of a hurtling pasta truck.

Noting the projected impact point, a television evangelist claimed divine retribution for the sins of Africa. "It's an African problem," he boomed. "Let the Africans fix it." Viewers barraged Congress with demands to halt funding for planetary defence.

Hamas stole an interceptor rocket and launched it sans warhead at Tel Aviv, killing 12 people. Twenty-seven minutes later AIPAC objected to installation of the rockets anywhere in the Middle East except Israel. The intercept programme slowed for five weeks to qualify alternative launch sites.

China announced the execution of ten corporate chieftains who had profiteered in rice, soya beans, petroleum and pirated movies. Tunnel-pocked Superfund sites sold at higher prices than Tokyo office lots. Desperate homeowners screamed and tussled to claim the best spots in abandoned mines. "*Fraternité*," pleaded the French president. The French parliament approved construction of a Riviera refuge for the president and his wife. A furore erupted following disclosure of the cost of the refuge's wardrobe closets and vanity mirrors.

Deep in the tropical forest 600 kilometres northeast of Kinshasa, a female bonobo named Lucy pouted her lips at her companion and scratched at a flea bite. Like most other organisms on the planet, *Pan paniscus* Lucy rarely made plans extending beyond her next meal or liaison. If she had been capable of understanding this statement and of formulating her response in sentence form, she would have demanded: "So what's your problem?"

Tests of a revolutionary technology showed promise. Money, people and equipment cascaded into crash development of a quantum resonance projector that would zap Goliath into a cloud of dust. On Day 769 this project, too, stopped dead. A patent troll had challenged the project IP. There'd been an errant MTA, or perhaps failure to file a CIP. Did the project have FTO? The US president suspended the dispute on the grounds of national security. After a delay of only six weeks the project resumed, though regrettably the weapon could no longer be ready before impact.

Day 824. In the spaceship at L5 a soft chime sounded the Point of Inevitability. Probability of imminent extinction of most Earth species now exceeded nine sigma. Preparing to depart the post they had held since the Trinity fission test, the two observers emitted soft bursts of methane from frustration and dyspepsia.

<* Confirm sequencing of Earth's gaiatype, ditto on-board storage of keystone biota breeding stock. *>

[+ Confirmed. I'll place the adverts. Maybe someone will give them a new home. +]

<* I hope so. I hate it when they have to be put down. You know, this job is stressing me out. I don't think I can stand to watch yet another intelligent species persuade itself of its mysterious immunity to natural selection. *>

[+ There you go again, attacking my religion. I wish you'd stop that. +] ■

Bruce W. Ferguson is a founder of two technology companies and the former editor of a humour magazine and of a law journal.